From the DEPARTMENT OF MOLECULAR MEDICINE AND SURGERY
Karolinska Institutet, Stockholm, Sweden

STRUCTURAL GENOMIC VARIATION IN HUMAN DISEASE

Maria Pettersson

Karolinska Institutet

Stockholm 2019

Cover art by Hannah Schwartz
Other illustrations and figures by Maria Pettersson et al.

Published by Karolinska Institutet

ISBN 978-91-7831-407-2
Printed by Eprint AB 2019

STRUCTURAL GENOMIC VARIATION

IN HUMAN DISEASE

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Maria Pettersson

Principal Supervisor:
Associate professor Anna Lindstrand
Karolinska Institutet
Department of Molecular Medicine and Surgery

Co-supervisors:
Dr Daniel Nilsson
Karolinska Institutet
Department of Molecular Medicine and Surgery

Professor Magnus Nordenskjöld
Karolinska Institutet
Department of Molecular Medicine and Surgery

Professor Elisabeth Syk Lundberg
Karolinska Institutet
Department of Molecular Medicine and Surgery

Dr Johanna Lundin
Karolinska Institutet
Department of Molecular Medicine and Surgery

Opponent:
Dr Nicole de Leeuw
Radboud University Nijmegen Medical Centre
Department of Human Genetics

Examination Board:
Associate professor Catharina Lavebratt
Karolinska Institutet
Department of Molecular Medicine and Surgery

Professor Marie-Louise Bondeson
Uppsala University
Department of Immunology, Genetics and Pathology

Associate professor Catarina Lundin
Lund University
Department of Laboratory Medicine
Division of Clinical Genetics

Deep in the forest, there is an unexpected clearing
that can only be reached by someone who has lost his way

Tomas Tranströmer

ABSTRACT

Structural variants (SVs) are physical changes in the structure of chromosomes and include both unbalanced copy number variants (CNVs) and balanced events (translocations, inversions and insertions). Many SVs constitute benign background variation and are found frequently in healthy individuals. Others may cause disease through gene disruption, deletion or duplication of dosage sensitive genes, or by disrupting the 3D structure of the genome.

In this thesis, we have delineated the exact structure of rearranged chromosomes and performed breakpoint junction analysis to study mutational signatures and underlying mechanisms of formation.

In **paper I**, we characterized and analyzed breakpoint junction sequences of 23 cytogenetically balanced translocations with mate-pair whole genome sequencing (WGS) and 17% of the translocations had microhomology and/or templated insertions in the breakpoint junctions, indicative of replication-based repair mechanisms. Genes were disrupted in 48% of breakpoints, highlighting a number of novel candidate genes and providing a molecular diagnosis in three cases. In **paper II**, we used targeted array comparative genomic hybridization and WGS to show that intragenic exonic duplications, formed through *Alu-Alu* fusion events, within *MATN3* and *IFT81* cause monogenic skeletal dysplasia disorders. Follow-up studies in primary cells and in zebrafish embryos showed that expression of a shorter IFT81 transcript alone is compatible with life. In **paper III**, we used WGS to investigate a benign complex chromosome rearrangement on chromosome 5p, detected in a healthy woman, which through unequal crossing-over during meiosis evolved into a pathogenic rearrangement including a duplication of the *NIPBL* gene in her daughter. In **paper IV**, we characterized the breakpoint junctions in 16 cytogenetically detected inversions. Contrary to what was expected, the vast majority of the resolved inversions were not mediated by inverted repeats through non-allelic homologous recombination. The mutational signatures in all the resolved inversions (11/16) indicate other mechanisms than ectopic recombination including replicative mechanisms in 2 cases. In **paper V**, we utilized WGS to perform a detailed characterization of 21 cases harboring multiple CNVs clustering on the same chromosome. The analysis revealed that multiple cellular mechanisms are involved in the formation of such SVs.

In conclusion, the results of this thesis show that WGS is a powerful way to delineate the structure of balanced, unbalanced and complex SVs. These studies have identified disease-causing aberrations, new candidate genes for further studies of neurodevelopmental disorders, and contributed to the understanding of how, when and why SVs arise.

LIST OF SCIENTIFIC PAPERS

I. Nilsson D*, Pettersson M*, Gustavsson P, Forster A, Hofmeister W, Wincent J, Zachariadis V, Anderlid BM, Nordgren A, Makitie O, Wirta V, Kaller M, Vezzi F, Lupski JR, Nordenskjold M, Syk Lundberg E, Carvalho CM, Lindstrand A. Whole-Genome Sequencing of Cytogenetically Balanced Chromosome Translocations Identifies Potentially Pathological Gene Disruptions and Highlights the Importance of Microhomology in the Mechanism of Formation. *Human mutation* (2017) 38(2):180-192

II. Pettersson M*, Vaz R*, Hammarsjo A, Eisfeldt J, Carvalho CMB, Hofmeister W, Tham E, Horemuzova E, Voss U, Nishimura G, Klintberg B, Nordgren A, Nilsson D, Grigelioniene G, Lindstrand A. *Alu-Alu* mediated intragenic duplications in *IFT81* and *MATN3* are associated with skeletal dysplasias. *Human mutation* (2018) 39(10):1456-1467

III. Pettersson M*, Eisfeldt J*, Syk Lundberg E, Lundin J, Lindstrand A. Flanking complex copy number variants in the same family formed through unequal crossing-over during meiosis. *Mutation Research* (2018) 812:1-4

IV. Pettersson M, Grochowski CM, Wincent J, Eisfeldt J, Cheung SW, Krepischi ACV, Rosenberg C, Lupski JR, Ottosson J, Lovmar L, Gacic J, Syk Lundberg E, Nilsson D, Carvalho CMB, Lindstrand A. Cytogenetically detected inversions are rarely formed by ectopic recombination between inverted repeats. *Manuscript*

V. Nazaryan-Petersen L*, Eisfeldt J*, Pettersson M, Lundin J, Nilsson D, Wincent J, Lieden A, Lovmar L, Ottosson J, Gacic J, Makitie O, Nordgren N, Vezzi F, Wirta V, Kaller M, Duelund Hjortshøj T, Jespersgaard C, Houssari R, Pignata L, Bak M, Tommerup N, Syk Lundberg E, Tümer Z, Lindstrand A. Replicative and non-replicative mechanisms in the formation of clustered CNVs are indicated by whole genome characterization. *PLoS Genetics* (2018) 14(11):e1007780

*Equal contribution

LIST OF RELATED SCIENTIFIC PAPERS

I. Viljakainen H, Andersson-Assarsson JC, Armenio M, Pekkinen M, Pettersson M, Valta H, Lipsanen-Nyman M, Makitie O, Lindstrand A. Low Copy Number of the *AMY1* Locus Is Associated with Early-Onset Female Obesity in Finland. *PLoS One* (2015) 10(7):e0131883

II. Bramswig NC, Ludecke HJ, Pettersson M, Albrecht B, Bernier RA, Cremer K, Eichler EE, Falkenstein D, Gerdts J, Jansen S, Kuechler A, Kvarnung M, Lindstrand A, Nilsson D, Nordgren A, Pfundt R, Spruijt L, Surowy HM, de Vries BB, Wieland T, Engels H, Strom TM, Kleefstra T, Wieczorek D. Identification of new *TRIP12* variants and detailed clinical evaluation of individuals with non-syndromic intellectual disability with or without autism. *Human genetics* (2017) 136(2):179-192

III. Pettersson M, Bergendal B, Norderyd J, Nilsson D, Anderlid BM, Nordgren A, Lindstrand A. Further evidence for specific *IFIH1* mutation as a cause of Singleton-Merten syndrome with phenotypic heterogeneity. *American journal of medical genetics Part A* (2017) 173(5):1396-1399

IV. Pettersson M*, Viljakainen H*, Loid P, Mustila T, Pekkinen M, Armenio M, Andersson-Assarsson JC, Makitie O, Lindstrand A. Copy Number Variants Are Enriched in Individuals With Early-Onset Obesity and Highlight Novel Pathogenic Pathways. *Journal of Clinical Endocrinology and Metabolism* (2017) 102(8):3029-3039

V. Hammarsjo A*, Wang Z*, Vaz R, Taylan F, Sedghi M, Girisha KM, Chitayat D, Neethukrishna K, Shannon P, Godoy R, Gowrishankar K, Lindstrand A, Nasiri J, Baktashian M, Newton PT, Guo L, Hofmeister W, Pettersson M, Chagin AS, Nishimura G, Yan L, Matsumoto N, Nordgren A, Miyake N, Grigelioniene G, Ikegawa S. Novel *KIAA0753* mutations extend the phenotype of skeletal ciliopathies. *Scientific reports* (2017) 7(1):15585

VI. Hofmeister W, Pettersson M, Kurtoglu D, Armenio M, Eisfeldt J, Papadogiannakis N, Gustavsson P, Lindstrand A. Targeted copy number screening highlights an intragenic deletion of *WDR63* as the likely cause of human occipital encephalocele and abnormal CNS development in zebrafish. *Human mutation* (2018) 39(4):495-505

VII. Costantini A, Skarp S, Kampe A, Makitie RE, Pettersson M, Mannikko M, Jiao H, Taylan F, Lindstrand A, Makitie O. Rare Copy Number Variants in Array-Based Comparative Genomic Hybridization in Early-Onset Skeletal Fragility. *Frontiers in Endocrinology* (2018) 9:380

VIII. Salehi Karlslatt K*, Pettersson M*, Jantti N, Szafranski P, Wester T, Husberg B, Ullberg U, Stankiewicz P, Nordgren A, Lundin J, Lindstrand A, Nordenskjold A. Rare copy number variants contribute pathogenic alleles in patients with intestinal malrotation. *Molecular Genetics & Genomic Medicine* (2018) 7(3):e549

IX. Eisfeldt J*, Pettersson M*, Vezzi F, Wincent J, Kaller M, Gruselius J, Nilsson D, Syk Lundberg E, Carvalho CMB, Lindstrand A. Comprehensive structural variation genome map of individuals carrying complex chromosomal rearrangements. *PLoS Genetics* (2019) 15(29):e1007858

*Equal contribution

CONTENTS

LIST OF ABBREVIATIONS

aCGH	Array comparative genomic hybridization
BCA	Balanced chromosomal aberration
bp	Basepair
BWA	Burrows-Wheeler aligner
CCR	Complex chromosomal rearrangement
CNV	Copy number variant
DNA	Deoxyribonucleic acid
FISH	Fluorescence *in situ* hybridization
FoSTeS	Fork stalling and template switching
IGV	Integrative Genomics Viewer
Indel	Insertion/deletion
kb	Kilobase (1,000 basepairs)
LCR	Low copy repeat
Mb	Megabase (1,000,000 basepairs)
MIM	Mendelian Inheritance in Man
MMBIR	Microhomology-mediated break-induced replication
MP	Mate-pair
NAHR	Non-allelic homologous recombination
NHEJ	Non-homologous end-joining
OMIM	Online Mendelian Inheritance in Man
PCR	Polymerase Chain Reaction
PE	Paired-end
RBM	Replication-based mechanism
SNV	Single nucleotide variant
SV	Structural variant
TAD	Topologically associated domain
WGS	Whole genome sequencing

1 INTRODUCTION

Structural variants (SVs) are physical changes in the structure of one or several chromosomes and can either be balanced (translocations, inversions, insertions) or unbalanced (copy number variants (CNVs); deletions, duplications). SVs may cause human disease by the disruption of genes or through copy number changes of dosage sensitive genes. In addition, structural variation of the genome can, apart from directly affecting specific genes, also change the genomic architecture and 3D structure of the DNA molecules, indirectly affecting genes by moving regulatory elements such as enhancers, promoters and silencers to a new location closer to or further away from target genes.

SVs implicated in human disease can be unique (found only in a single individual/family with unique breakpoints) or recurrent (repeatedly formed with similar breakpoints in unrelated individuals). The mechanisms underlying the formation of recurrent and non-recurrent SVs differ; recurrent SVs are most commonly formed through errors during homologous recombination and non-recurrent SVs arise through DNA replication errors or mistakes during repair of double strand breaks in the DNA molecules (Carvalho and Lupski, 2016; Currall, et al., 2013). Some chromosomal regions are prone to recurrent rearrangements due to the local genomic architecture and diseases caused by such genomic rearrangements are generally referred to as genomic disorders (Lupski, 1998).

In 1959 the first genetic aberration was identified; Trisomy 21, or Down syndrome, caused by a whole extra chromosome 21. This aberration is visible through a microscope and SVs involving smaller parts of chromosomes can also be microscopically visible through G-banding of chromosomes, if the genomic segments involved are larger than ~5-10 Mb. Karyotyping with G-banding of chromosomes was introduced in the 1970s (Yunis and Sanchez, 1973), and is still commonly used in clinical practice. Fluorescence *in situ* hybridization (FISH) was introduced in the mid 1980s with fluorescent probes hybridizing to specific chromosomes or specific parts of a chromosome, and the genome resolution was improved to 100 kb - 1 Mb (Pinkel, et al., 1986). In mid 2000s, the microarray technique was introduced, and especially the array comparative genomic hybridization (aCGH) (Tchinda and Lee, 2006) was revolutionizing the clinical practice of detecting SVs, with a resolution of ~50 kb, depending on amount and distribution of probes. However, aCGH only detects unbalanced SVs. Whole-genome sequencing (WGS) was introduced around 2010 (Metzker, 2010) and has rapidly developed into a crucial tool for detailed characterization of SVs. As

for now, the massive amount of data from WGS is limiting the use for WGS as a screening method for SVs, and it is currently mostly used to characterize SVs that have been detected through other methods. However, the tools for interpreting WGS data are currently being developed and tested for "WGS first" use.

1.1 STRUCTURAL VARIATION IN THE HUMAN GENOME

1.1.1 Normal structural genomic variation

The human genome consists of approximately 3 billion DNA base pairs, of which less than 2% consists of exons that codes for proteins. The remaining >98% is made up of regulatory sequences, repeat elements and pseudogenes, as well as non-coding genetic material that to a large extent still is of unknown function. This part of the genome was referred to as "junk DNA" for a long time, but the more the non-coding part of the genome is studied, the more interesting it gets and it is increasingly obvious that "junk DNA" is absolutely crucial for cells to function properly. A recent example is the multinational project called Synthetic Yeast 2.0, where the main purpose is to synthesize all the yeast chromosomes from scratch (Mitchell, et al., 2017). Some parts of the genome may be removed without obvious effects, while other parts of the non-coding DNA need to be intact for the yeast to survive; for example, deleting the subtelomeric DNA severely affected gene expression and caused silencing of genes that were not supposed to be silenced (Mitchell, et al., 2017).

Ever since the human genome was sequenced and published (International Human Genome Sequencing, 2004), it has been clear that there is an enormous amount of normal variation among individuals (Genomes Project, et al., 2015). In the 1000 Genomes Project, the genomes of 2,504 individuals from 26 populations were reconstructed and found that a typical genome differed from the reference human genome at 4.1-5 million sites and structural variants affected ~20 million base pairs of sequence (Genomes Project, et al., 2015). The specific normal variants in the genome vary between populations, and it is therefore important to sequence large numbers of individuals from as many populations as possible. On average, each individual carries ~25 variants in genes previously implicated in human disease (Genomes Project, et al., 2015) and the available databases of normal variants are crucial for filtering out the disease-causing variants from variants without clinical impact.

Short-read WGS allows for the detection of single nucleotide variants (SNVs), insertions/deletions (indels) and SVs (balanced and unbalanced) in a single experiment. The introduction of WGS has speeded up the amount of human genomes sequenced and hence the detection of normal background variation. Today, we know that some genes can be completely knocked out without an apparent phenotypic effect, while others are highly sensitive to variation (Sudmant, et al., 2015).

1.1.2 Role of repeat elements in structural genomic variation

Over 50% of the human genome consists of repetitive sequences, so called repeat elements, that are present in multiple copies throughout the genomic sequence (de Koning, et al., 2011). Repeat elements are commonly divided into subgroups based on their structure and characteristics, for example tandem repeats (e.g. microsatellites, telomeres), interspersed repeats (e.g. mobile elements, pseudogenes) and low copy repeats (LCRs, or segmental duplications). We know the function or part of the function of some of these repeat elements, such as the sequences that comprise the telomeres that are crucial for chromosome dynamics during cell replication (Blackburn, 1991), or mobile elements that are widely recognized as drivers of genetic evolution (Kazazian, 2004).

However, we also know that some repeat elements are the underlying cause of recurrent structural genomic rearrangements, of which some cause disease. A common example of how repeat elements predispose to disease-causing structural variants is the 17p deletions and duplications. The proximal p-arm of chromosome 17 is both gene- and LCR-rich and has been described in numerous cases with different rearrangements in constitutional and cancer-associated chromosomal aberrations (Barbouti, et al., 2004; Pentao, et al., 1992; Stankiewicz, et al., 2004). In a ~7.5 Mb region on 17p that was investigated in a paper by Stankiewicz et al., it was found that LCRs constituted over 23% of the genomic sequence, explaining why this particular part of the genome seems to be more unstable than other genomic regions (Stankiewicz, et al., 2003).

Unstable regions of the genome are still detected, and different types of repetitive elements seem to be responsible for, and predispose to, different types of rearrangements. Investigations of various diseases and genes have demonstrated the role of *Alu* elements in the formation of SVs (Boone, et al., 2011; Boone, et al., 2014; Gu, et al., 2015; Song, et al., 2018). *Alu* elements are repetitive sequences belonging to the primate-specific short interspersed nuclear element (SINE) family of mobile DNA elements. They comprise ~11%

of the human genome and are present in more than 1 million copies in a haploid genome (Lander, et al., 2001). In a large study including 65 families with uncharacterized ciliopathies, a small (6.7 kb) tandem duplication in *IFT140* was identified in eight families (Geoffroy, et al., 2018). It was found that the duplication was not a founder variant, but that the breakpoints overlapped distinct *Alu* elements with high sequence identity (68-81%) and resulted in an *Alu* hybrid on both sides of the duplication (Geoffroy, et al., 2018). In a recent study by Song et al., addressing the genomic instability caused by *Alu* elements, 12,074 OMIM genes were tested for the relative risk of *Alu-Alu* mediated rearrangements and 47 duplications, 40 deletions and two complex rearrangements were fine-mapped (Song, et al., 2018). It was found that 94% of the candidate breakpoints were at least partially mediated by *Alu* elements, further strengthening the role of *Alu* elements in gene and genome evolution as well as in mediating human disease (Song, et al., 2018).

1.2 STRUCTURAL VARIATION IN HUMAN DISEASE

1.2.1 Balanced chromosomal aberrations

The overall incidence of balanced chromosomal aberrations (BCAs), such as translocations (Figure 1A), inversions (Figure 1B-C) or insertions (Figure 1D), has been estimated in to be approximately 0.2% in an unselected newborn population (Jacobs, et al., 1992). In the same study, it was calculated that about 20% of translocations occur *de novo* with an estimated mutational rate of 2.7e-4 per gamete per generation (Jacobs, et al., 1992). Cytogenetic investigation of couples experiencing recurrent pregnancy loss is standardized and is the most common reason why BCAs in individuals with no other clinical phenotype are found (Jacobs, 1987). Due to a risk of malsegregation of the rearranged chromosomes and errors in meiotic recombination, BCA carriers have an increased risk of having children with unbalanced rearrangements causing severe disease.

It has been estimated that about 6% of *de novo* reciprocal balanced translocations are associated with a serious congenital anomaly, apparent before 1 year of age (Warburton, 1991). Even though a significantly higher prevalence of BCAs has been identified in cohorts with neurodevelopmental disorders (1.5%) (Funderburk, et al., 1977) and autism spectrum disorder (ASD, 1.3%) (Marshall, et al., 2008), there are no solid numbers on the amount of mildly affected (mild intellectual disability, ADHD, dyslexia, learning difficulties) balanced

translocation carriers. A recent study with long-term follow-up (mean follow-up time 17 years) showed that 26.8% of the *de novo* BCA carriers that were investigated presented with a clinical diagnosis, of which most were neurodevelopmental disorders (Halgren, et al., 2018). The study suggests that the risk for mild neurodevelopmental disorders that are not obvious within the first year of life in *de novo* BCA carriers could be 2-3-fold higher than the risk for severe congenital diseases or malformations (predicted to approximately 6%), and that more studies with long-term follow-up are needed (Halgren, et al., 2018). BCAs disrupting single genes may help pinpoint directly disease-causing loci (Bramswig, et al., 2017; Hofmeister, et al., 2015). In addition, it has been shown that recurrent inversions in especially unstable chromosomal regions cause no clinical phenotype in the carrier, but can increase the risk of microdeletions forming in the offspring, as in the case of Williams-Beuren syndrome (MIM:194050, 7q11.23) and Prader-Willi syndrome/Angelman syndrome (MIM:176270/MIM:105830, 15q11-q13) (Gimelli, et al., 2003; Osborne, et al., 2001).

Pinpointing the exact breakpoints is crucial for determining whether a balanced chromosomal aberration is involved in the clinical phenotype presented by the carrier. Historically, time consuming mapping of cytogenetically balanced chromosomal aberrations using pulse-field gel electrophoresis was a successful approach to new gene discovery with one prominent example being *NF1* (causing neurofibromatosis type 1, MIM:162200) (Fountain, et al., 1989). With use of WGS, breakpoint mapping and disease gene discovery has accelerated (Chiang, et al., 2012; Finelli, et al., 2007; Fruhmesser, et al., 2013; Fukushi, et al., 2018; Talkowski, et al., 2012).

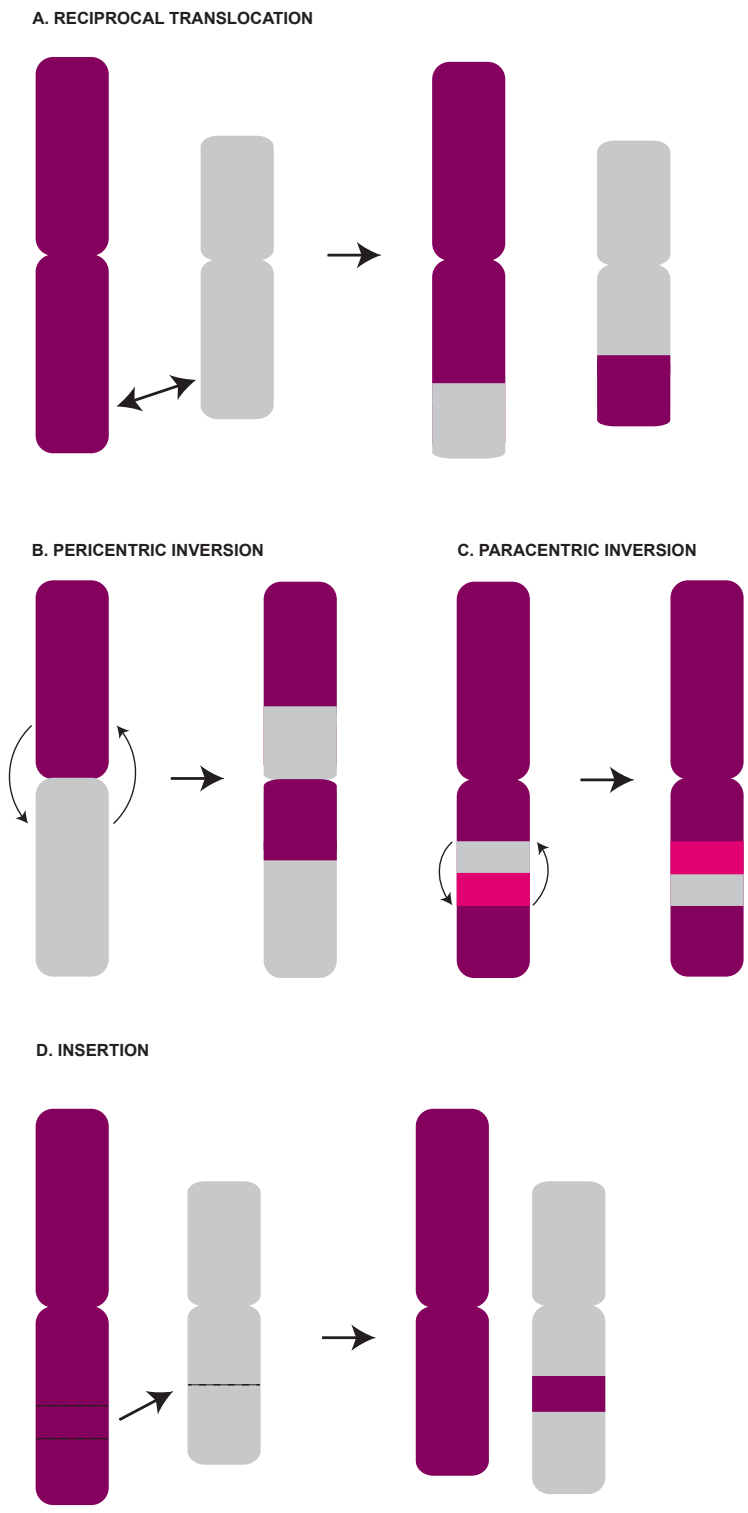

Figure 1. *Reciprocal balanced translocations (A), pericentric inversions (B), paracentric inversions (C) and insertions (D) are structural genomic variants commonly classified as balanced chromosomal aberrations. However, genes could be directly or indirectly affected by the breakpoints.*

1.2.2 Copy number variants

Copy number variants (CNVs; deletions and duplications (Figure 2)) have been implicated in numerous diseases, but are also known to be present in healthy human genomes as benign polymorphic variants, present in >1% of individuals in different populations (Zhang, et al., 2009a). Between 4.5%-9.5% of the genome has been estimated to contribute to copy number variation (Zarrei, et al., 2015) and it appears that as many as approximately 100 of the total protein-coding genes can be completely lost (homozygous deletion) without apparent phenotypic consequences (Zarrei, et al., 2015).

Many disease-causing CNVs are large, involving several megabases of nucleotides, such as Prader-Willi syndrome/Angelman syndrome (MIM:176270/105830, 15q11-q13 deletion), Williams-Beuren syndrome (MIM:194050, 7q11.23 deletion) and Potocki-Lupski syndrome (MIM:610883, 17p11.2 duplication). However, small intragenic CNVs have with increased quality of detection methods been more commonly reported as contributors or independent drivers of disease (Lieden, et al., 2014; Lindstrand, et al., 2016).

Conventional cytogenetic investigations mentioned in the introduction, such as karyotyping and FISH (fluorescent *in situ* hybridization) mapping, can only detect gain or loss of hundreds of kilobases (kb) (FISH) or megabases (Mb) (karyotyping) of nucleotides (Warburton, 1980). Today, for whole-genome investigation of copy number variation, chromosomal microarrays such as SNP arrays or comparative genomic hybridization arrays (aCGH) are the most commonly used methods. Regular aCGH platforms used in clinical practice, with evenly spread probes across the entire genome, commonly has a resolution of approximately 25-50 kb, depending on the number and distribution of probes. However, targeted arrays may increase the resolution down to a few hundred bases in specific target regions (Lindstrand, et al., 2016).

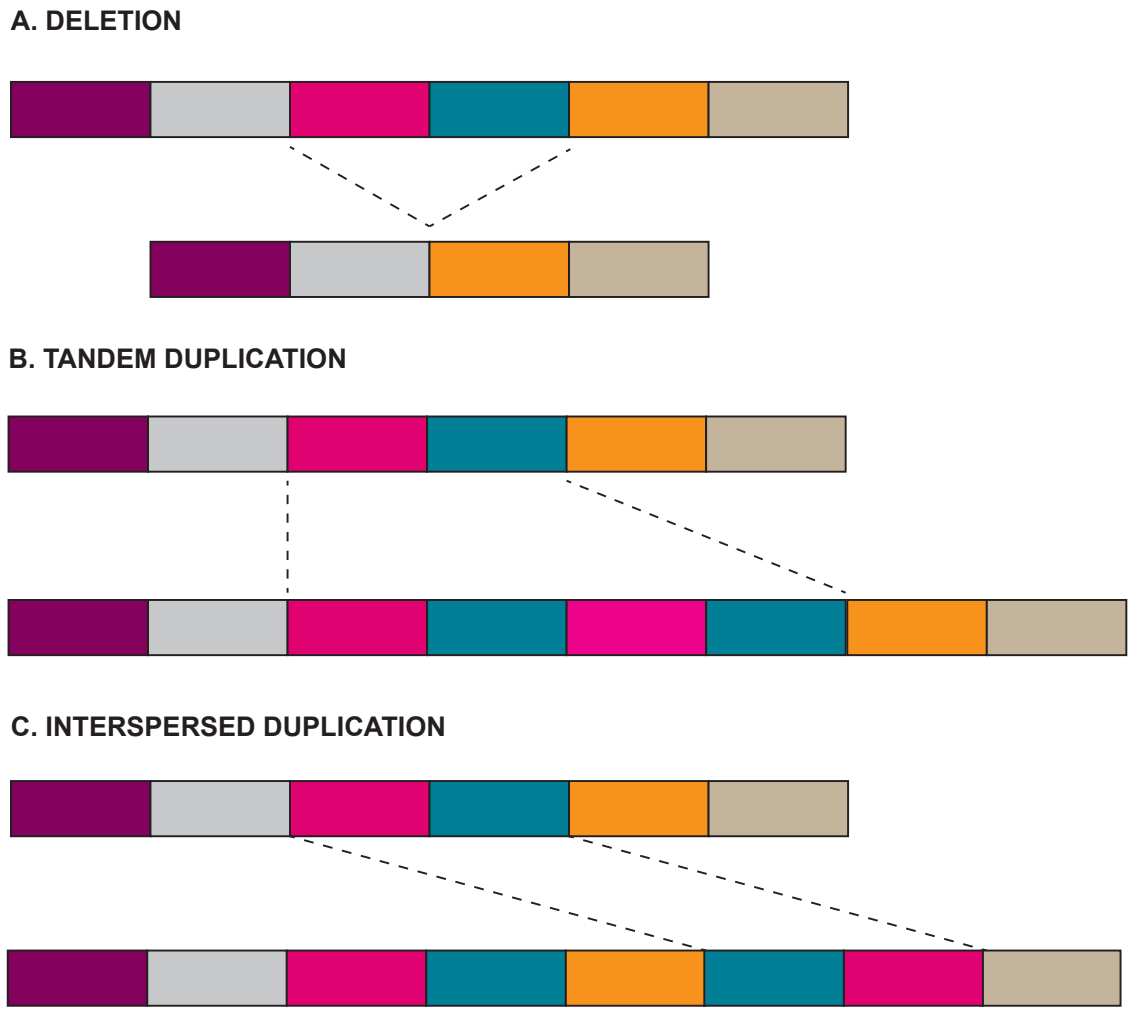

Figure 2. *Copy number variants can be either deletions (A), where a piece of genetic material is completely lost, or duplications, where a piece of genetic material either has been duplicated in tandem (B), or inserted into another genomic position anywhere in the genome, sometimes in inverted orientation (C).*

1.2.3 Complex chromosomal rearrangements

Complex chromosomal rearrangements (CCRs) involve a single or several chromosomes with three or more breakpoints as detected by chromosome analysis (Liu, et al., 2012). Complex genomic rearrangements (CGRs) are defined in the same way but also include submicroscopic rearrangements that are not seen on chromosome analysis but only detected using molecular tools, such as aCGH or WGS. Complex rearrangements of chromosomes have traditionally been detected using conventional cytogenetic methods such as karyotyping, FISH and aCGH. However, the poor resolution of these methods has made it difficult to fully

characterize the genomic complexities. Using WGS with paired or linked reads, a number of CCRs have now been solved down to nucleotide level and commonly, additional complexity is revealed by this high-resolution analysis (Aristidou, et al., 2018). More than half of *de novo* CCRs are associated with a clinical phenotype (Madan, et al., 1997), and as with regular CNVs the clinical symptoms may be caused by I) direct disruption of disease-associated genes in the breakpoints, II) deletions or duplications of dosage sensitive genes, or III) physical changes in the 3D structure of the genome (Hodge, et al., 2014; Lupianez, et al., 2015; Lupianez, et al., 2016; Schluth-Bolard, et al., 2013; Talkowski, et al., 2012). Small (from 1 bp to a few kb) imbalances in the breakpoint junctions are generally only detected using sequencing techniques, which additionally can pinpoint the underlying mechanism of formation (see 1.4.3 Complex rearrangements).

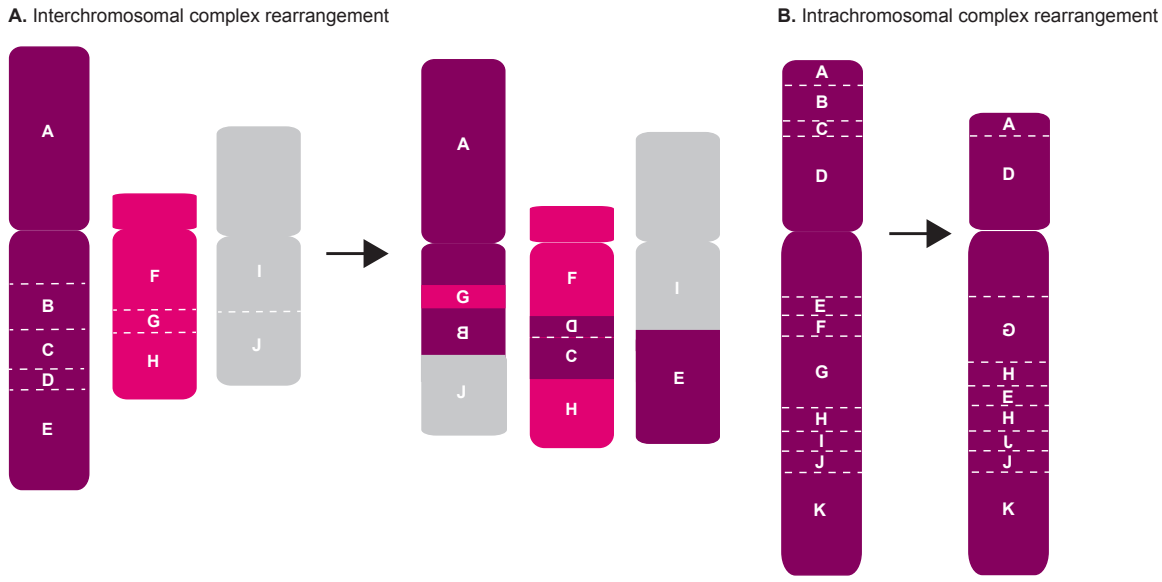

Figure 3. *Complex chromosomal rearrangements can involve several chromosomes (**A**, interchromosomal) or a single chromosome (**B**, intrachromosomal). In example A, three chromosomes are involved in the rearrangement and during the reassembly segments B and D have been inverted. In example B, a single chromosome has been shattered and during reassembly, segments B, C, F and I have been lost, segment G has been inverted and segments H and J have been duplicated, with segment J also being inverted.*

1.2.4 Structural variation of the non-coding human genome

It has been known for a long time that long stretches of non-coding sequences, sometimes referred to as gene deserts, can contain regulatory elements important for gene expression. Although these regulatory elements can be located megabases away from the actual gene, a disruption of a seemingly harmless part of the genome was already in 2002 shown to be the cause of disease due to dysfunctional gene expression (Lettice, et al., 2002; Nobrega, et al., 2003). The impact of structural variants on human disease was understood on a deeper level when position effects and the enhancer adoption concepts were reported, and new cases could be solved when the 3D structure of the genome was investigated (Lettice, et al., 2011). One of the proposed mechanisms to explain this phenomenon has been topologically associated domains (TADs). TADs are highly conserved genomic segments, megabases in size, which divide the genome into units with a high degree of intra-domain interaction, separated from one another by topological boundary regions, blocking the interaction between neighboring TADs. TADs are formed by chromatin loops, often involving "looping factors" such as the CCCTC-binding factor (CTCF) and cohesin, and contribute to the healthy genome by preventing faulty activation of genes. Structural variants, although cytogenetically balanced and not disrupting any coding sequences, may disrupt TADs and move enhancers, causing one or several genes to be up- or downregulated without actually interfering with the gene itself (Krijger and de Laat, 2016; Ordulu, et al., 2016; Redin, et al., 2017). As mentioned previously, seemingly balanced chromosomal rearrangements are more common in patients with various disease phenotypes. With this in mind, and the fact that more than 90% of BCA breakpoints are located in non-coding DNA, this is likely to be an important disease-causing mechanism (Krijger and de Laat, 2016; Lettice, et al., 2002; Maurano, et al., 2012; Nobrega, et al., 2003; Schaub, et al., 2012).

1.3 WHOLE GENOME SEQUENCING FOR DETECTION OF STRUCTURAL GENOMIC VARIATION

The 1000 Genomes Project started reporting structural variant data generated from aCGH, PCR and SNP arrays in 2010, focusing on CNVs (Mills, et al., 2011; Pennisi, 2010). In 2015, the first large report of structural variants from WGS data was published, which included inversions, mobile elements and very small SVs (<1 kb) (Sudmant, et al., 2015). In comparison to aCGH and FISH, WGS allows for simultaneous detection of single nucleotide

variants (SNVs), CNVs and BCAs. Since the introduction of massive parallel sequencing (MPS), the data has mostly been used to detect SNVs, and among the first MPS implementations were exome sequencing, exclusively used for SNV detection. Today, we know that WGS can detect most SVs, and provides detailed information such as position of segments, mutational signatures of junctions and exact location of breakpoints. The currently most commonly used human WGS protocol (Illumina Paired-End (PE) short-read WGS) provides the orientation of the two paired reads; an abnormal orientation may indicate an inversion, or in case of a duplication it shows if the extra segments is inserted or in tandem. Short insert sizes of PE WGS libraries often provide nucleotide resolution of breakpoints due to reads spanning the actual breakpoint, so called split reads (Mills, et al., 2011). The second most commonly used WGS protocol is Mate-Pair (MP) sequencing and PE and MP WGS protocols are very similar but works optimally for different types of SVs. In PE WGS, short fragments (300-800 bp) of DNA are sequenced while MP libraries use longer DNA fragments (1,000-4,000bp) that are first circularized, the sequence pairs are therefore oriented in inverse orientation and are generally further apart than in PE libraries. Generated reads for both methods are commonly 100-150 base pairs in length. The raw WGS data is aligned to the human reference genome, and many different bioinformatics tools are used to identify and annotate variants. Detection of SVs from either PE or MP sequencing libraries is based on read depth, discordant sequence pairs mapped to unexpected places or directions (Fullwood, et al., 2009) and *de novo* assembly. Examples of callers for structural variant detection are CNVnator (Abyzov, et al., 2011), Manta (Chen, et al., 2016), and TIDDIT (Eisfeldt, et al., 2017).

A limitation for precise mapping of breakpoint junctions are low copy repeats (LCRs), common in recurrent structural variants, and regular short-read libraries such as PE and MP will often not sufficiently cover the entire repeat. For these variants, other sequencing methods have been developed, commonly referred to as third-generation sequencing or long-read sequencing; PacBio (Pacific Biosciences) and Nanopore (Oxford Nanopore Technologies) are examples of third-generation sequencing technologies. Other technologies utilize barcoded reads to synthesize long reads bioinformatically (10X Genomics Chromium) or optical maps to generate low-resolution maps of the genome (Bionano Genomics). PacBio sequencing captures sequence information during the replication process of a single molecule and generates reads around 10 kb in length (Rhoads and Au, 2015) while Nanopore sequencing generates sequence information by having single DNA strands passing through a biological pore and generates an output of 10-100 kb reads, depending on the fragment sizes of the input DNA (Lu, et al., 2016). Linked-read sequencing provided by 10X Genomics

Chromium protocol utilizes long DNA molecules that are barcoded and separated into droplets, and sequenced barcodes links the sequences that belong to the same long DNA molecule (Eisfeldt, et al., 2019). Bionano is not a sequencing technique but instead uses fluorescently tagged probes to image the genome and reveal structural changes but no sequence variation (English, et al., 2015).

1.4 STRUCTURAL VARIANT MECHANISMS OF FORMATION

1.4.1 Recombination mechanisms

Mutational signatures in the breakpoint junctions of genomic rearrangements may provide insight into the mechanism involved in the original rearrangement formation (Figure 4). For example, the presence of large segments of homologous DNA (>200 bp) flanking the junction would imply non-allelic homologous recombination (NAHR) as the chromosome break and repair mechanism (Stankiewicz and Lupski, 2002; Waldman and Liskay, 1988).

Figure 4. Mutational signatures in breakpoint junctions provide clues on the underlying mechanism of formation. *NHEJ; non-homologous end-joining, FoSTeS; fork-stalling and template-switching, SINE; short interspersed nuclear element, LINE; long interspersed nuclear element, SD; segmental duplication, NAHR; non-allelic homologous recombination, MMBIR; microhomology-mediated break-induced replication*

Several lines of evidence show that sequences with repetitive features are targets for NAHR, and are hence prone to cause recurrent balanced and unbalanced genomic rearrangements (Liu, et al., 2012). In recurrent structural variants, the breakpoints cluster in the same genomic intervals and the resulting aberration is similar in size and position between carriers (Liu, et al., 2012). NAHR was the first mechanism identified to be involved in chromosomal rearrangement formation. NAHR is also the mechanism responsible for genomic disorders (Lupski, 1998) where unequal crossing over between highly similar genomic segments on homologous chromosomes (i.e. LCRs or segmental duplications) in direct orientation result in deletions and duplications (Figure 5). Inverted repeats may result in inversions and crossing-over between LCRs in non-homologous chromosomes can produce recurrent translocations (Ou, et al., 2011). It has been proposed that 300-500 bp of perfectly matching DNA sequence is the minimal efficient processing segment for NAHR to occur (Reiter, et al., 1998).

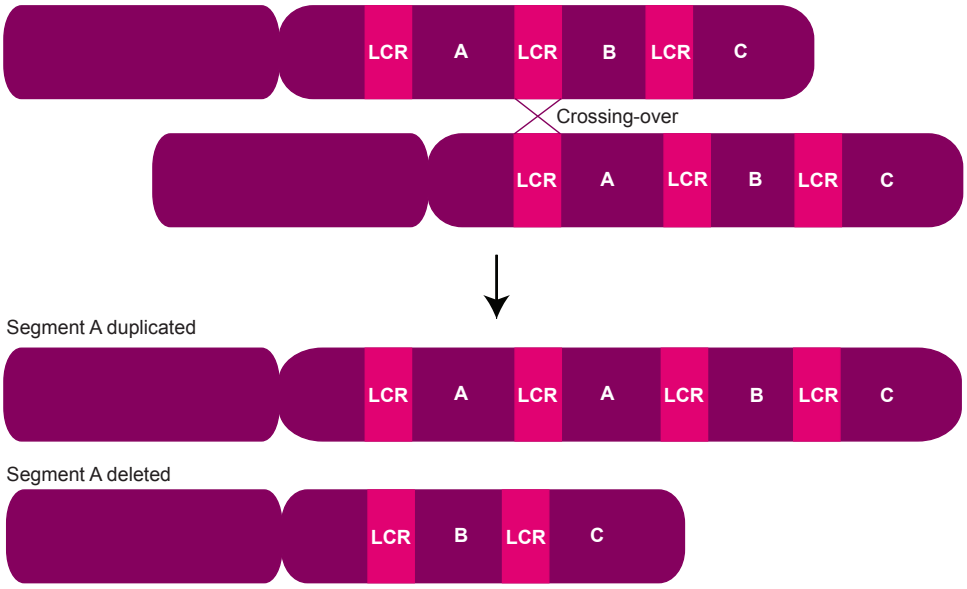

Figure 5. *Non-allelic homologous recombination (NAHR) can occur when low copy repeats (LCRs) on homologous chromosomes recombine due to sequence similarity. Some parts of the genome are more prone to NAHR-mediated rearrangements due to high LCR content in those areas. Recurrent structural variants are formed through NAHR.*

1.4.2 Replication mechanisms

Replication-based mechanisms (RBMs) such as fork stalling and template switching/microhomology mediated break-induced replication (FoSTeS/MMBIR) have been proposed as the underlying mechanism of formation of complex and non-recurrent structural variants in humans (Lee, et al., 2007; Zhang, et al., 2009b). Mutational signatures implying RBM-mediated rearrangements include presence of microhomology, small templated insertions and novel SNVs in close proximity of the breakpoint junction and duplications/triplications accompanying large inverted genomic segments (Carvalho, et al., 2011).

In Pelizaeus-Merzbacher disease (MIM:312080), caused by non-recurrent duplications or deletions of proteolipid protein 1 (*PLP1*), high-resolution aCGH and subsequent breakpoint sequence analyses have revealed interspersed stretches of DNA within the duplicated sequence and sequence complexity at the junctions (Lee, et al., 2007). These findings suggested that mechanisms involving errors of DNA replication, rather than the generally considered meiotic recombination mechanisms, could cause structural genomic rearrangements. Briefly, the FoSTeS model proposed was that during DNA replication, the active replication fork stalls and switches template using complementary template microhomology to prime re-annealing. The process may involve forks linearly far apart in the genome but in close 3D proximity (Lee, et al., 2007).

1.4.3 Complex rearrangements

Whole-genome sequencing studies have identified many CGRs that appear to be derived from a single catastrophic event, sometimes referred to as “chromosome pulverization” and what appears to be a random reassembly of the chromosome fragments. Two separate complex catastrophic phenomena are often described in congenital chromosomal rearrangements: chromothripsis and chromoanasynthesis (Liu, et al., 2011; Masset, et al., 2016).

Chromothripsis has been proposed to originate in a multi-step process, where chromosome segregation errors first result in formation of micronuclei, and chromosomes trapped within the micronuclei undergo delayed DNA replication and become fragmented. In the next cell cycle, the damaged chromatin undergoes complex rearrangements and subsequent fusion of the micronucleus with the main nucleus results in a daughter cell carrying a highly rearranged

chromosome. The proposed mechanism for the repair process is non-homologous end-joining (NHEJ) (Kloosterman, et al., 2012; Zhang, et al., 2015). The concept of micronuclei would explain the mystery of how such catastrophic cellular events could be isolated to relatively small parts of the genome. Although first described in cancer cells, chromothripsis is also a likely involved in germline complex *de novo* structural rearrangements (Kloosterman, et al., 2011).

Chromoanasynthesis is mediated by replications errors (Lee, et al., 2007). The first report on chromoanasynthesis came in 2011 when Liu et al. investigated patients with developmental anomalies and discovered several duplications, triplications and deletions, and subsequent sequencing of the breakpoint junctions revealed microhomology and templated insertions, consistent with FoSTeS/MMBIR (see 1.3.2 Replication mechanisms) (Liu, et al., 2011). A distinct form of chromoanasynthesis, referred to as atypical chromoanagenesis, was first reported in 2016 by Masset et al., who reported complex rearrangements containing only copy number gains (Masset, et al., 2016). The chromosomal aberrations were reported to only affect a single chromosome, and being dispersed throughout the entire chromosome. The main difference from the other type of chromoanasynthesis was the presence of duplications only and up to 40 bp of non-templated insertions in the breakpoint junctions, and the proposed joining mechanism was DNA polymerase Polθ-driven alternative NHEJ (Masset, et al., 2016).

2 AIM OF THESIS

The overall purpose of this thesis was I) to characterize human chromosomal aberrations such as CNVs, translocations, inversions, and complex rearrangements, in order to elucidate possible mechanisms of formation as well as mechanisms of disease, and II) improve the diagnostic methods for SV detection. We have performed detailed studies of rare chromosomal aberrations, in hope of identifying disease-causing genetic alterations.

By using whole genome sequencing, we have a unique opportunity to identify and characterize the breakpoints of structural variants with nucleotide resolution, which is essential for deeper understanding of the underlying mechanisms of formation, as well as pinpointing of specific genes in or in close proximity of the breakpoint(s).

Specifically, the aims of this thesis were to:

I. Pinpoint and characterize breakpoint junctions of structural variants at the nucleotide level

II. Use mutational signatures to elucidate possible mechanisms of formation

III. Identify known or novel disease genes in or in close proximity of the breakpoints

IV. Identify the cellular process that caused the chromosomal aberration to occur

3 MATERIAL AND METHODS

3.1 PATIENTS AND CLINICAL DATA

All patients included in **paper I** were collected at the Clinical Genetics department, Karolinska University Hospital (Stockholm, Sweden), or in one case at Helsinki University Hospital (Helsinki, Finland). The study covered 22 individuals, in which conventional chromosome analysis had identified at least one cytogenetically balanced translocation (one patient was a carrier of two separate translocations). Twelve of the patients had been referred for chromosome analysis because of an affected phenotype and ten had first been referred for amniocentesis because of age, worry, ultrasound findings, multiple previous miscarriages, or previous birth of a child with an unbalanced karyotype. Eight of these individuals were unaffected, and two had mild neurodevelopmental disorders.

Both patients in **paper II** were first identified at the Clinical Genetics department, Karolinska University Hospital (Stockholm, Sweden) and were originally referred for genetic investigation because of clinically affected phenotypes (skeletal dysplasia).

The proband in **paper III** was referred to the Clinical Genetics department, Karolinska University Hospital (Stockholm, Sweden) due to clinically affected phenotype presenting within the first year of life. Clinical aCGH revealed a large duplication and segregation analysis of the parents revealed that the mother carried two small deletions, flanking the area that was duplicated in the proband, which prompted for further analysis in a research setting.

All patients included in **paper IV** were collected at the Clinical Genetics department, Karolinska University Hospital (Stockholm, Sweden), Linköping University Hospital (Linköping, Sweden), Sahlgrenska University Hospital (Gothenburg, Sweden), Baylor College of Medicine (Houston, Texas, USA) and University of São Paolo (São Paolo, Brazil). A total of 24 patients were included, carrying 16 unique cytogenetically visible inversions (13 pericentric and three paracentric). All patients had originally been referred for chromosomal analysis due to a clinically affected phenotype, fertility problems or a clinically affected child/intrauterine fetal death.

In **paper V**, we reported 21 individuals with clustered CNVs. Out of the 21 included individuals, five had been collected at the Kennedy Center (Rigshospitalet, Copenhagen, Denmark), two at Sahlgrenska University Hospital (Gothenburg, Sweden), one at Linköping University Hospital (Linköping, Sweden) and 13 at the Karolinska University Hospital

(Stockholm, Sweden). All included cases had been referred to each medical center for genetic investigation using aCGH due to congenital developmental disorders, intellectual disability or autism.

3.2 MOLECULAR ANALYSES

3.2.1 Cytogenetic analyses

3.2.1.1 Karyotyping

Metaphase slides were prepared from peripheral blood cultures, according to standard protocols. Chromosome analysis (Figure 6) was performed after G-banding with an approximate resolution of 550 bands per haploid genome. At least 10 metaphases were analyzed for each patient.

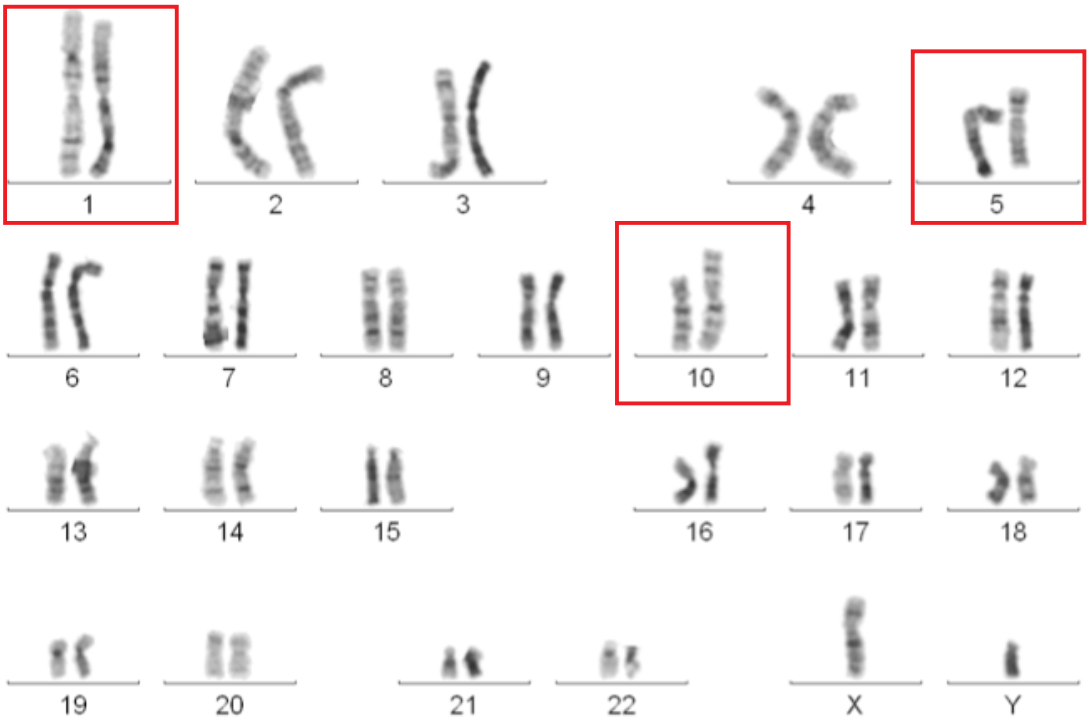

Figure 6. *Karyotype visualizing a complex chromosomal rearrangement involving chromosome 1, chromosome 5 and chromosome 10 (46,XY,t(1;10;5)(q32;p12;q31). Larger chromosomal rearrangements involving pieces >10 Mb can be visualized using traditional G-banding chromosome analysis. The patient is reported in (Lindstrand, et al., 2010) and (Eisfeldt, et al., 2019).*

3.2.1.2 Array comparative genomic hybridization (aCGH)

The aCGH studies were performed using three different microarray designs: I) a clinically used 180K design with 180,000 probes evenly covering the genome, II) a 1M medical exome array with 1,000,000 probes targeting 6,000 known disease-associated genes, and III) a custom design, designed using eArray, an online web tool from Agilent Technologies (Palo Alto, CA, USA). The custom design was an Agilent 2x400K high-definition comparative genomic hybridization microarray design (Agilent Technologies), consisting of 400,000 oligonucleotide probes. Out of the 400,000 probes, approximately 180,000 probes evenly covered the genome with a resolution of approximately 25-50 kb. Remaining probes targeted 1989 genes found in the cilia proteome and/or involved in malformation syndromes and intellectual disability. The resolution within targeted genes was 1 probe per 100 bp in coding sequences, and 1 probe per 500 bp in noncoding sequences. The array slides used for the experiments were ordered from Oxford Gene Technology (OGT, Oxfordshire, UK).

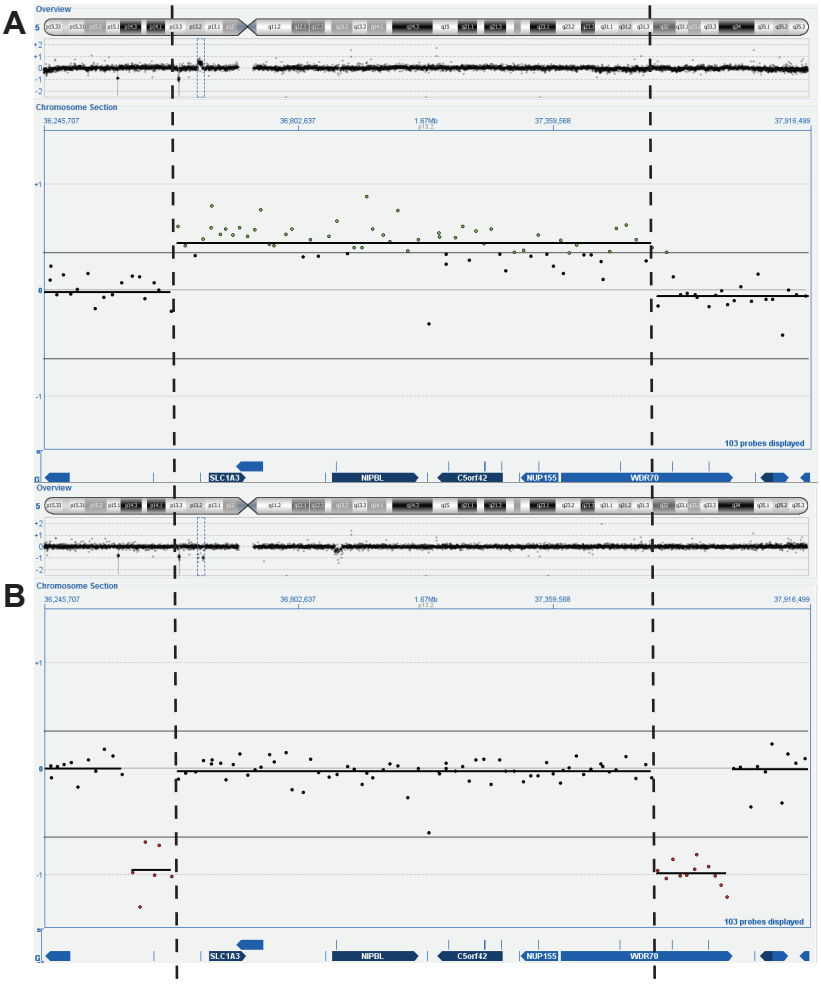

Figure 7. Array comparative genomic hybridization (aCGH) data visualized in CytoSure Interpret Software. *A large duplication was identified in a patient with developmental delay (A), and two small deletions flanking the duplication were found in the mother (B). Figure adapted and modified from* ***paper III****.*

3.2.2 Whole genome sequencing

3.2.2.1 30X PCR-free paired-end sequencing protocol

In **paper II**, **paper III, paper IV and paper V**, samples were subjected to WGS using a 30X PCR-free PE WGS protocol at National Genomics Infrastructure (NGI), Science for Life Laboratory, Stockholm, Sweden. PE-WGS protocols generate short (2x150 bp) paired reads, suitable for detection of SVs through both discordant read pairs and aberrant read depth. Data was processed using the NGI-piper and SVs were detected and analyzed using the in-house FindSV pipeline (https://github.com/J35P312/FindSV) that combines CNVnator V0.3.2 (Abyzov, et al., 2011) and TIDDIT (Eisfeldt, et al., 2017) and generates a single variant call format (VCF) file. The VCF file was annotated using variant effect predictor (VEP) and filtered based on the VCF file quality flag (McLaren, et al., 2016). Finally, the VCF file was sorted based on a local structural variant frequency database consisting of 351 patient samples. Split reads were identified when visualizing the breakpoint regions of each variant in IGV (http://software.broadinstitute.org/software/igv/) (Robinson, et al., 2011) and the exact position of the breakpoints could be determined in most cases using the BLAT (https://genome.ucsc.edu/cgi-bin/hgBlat) (Kent, 2002).

3.2.2.2 3X 2.5kb insert-size mate-pair sequencing protocol

In **paper I** and **paper V**, we used a low-coverage (3X) MP sequencing protocol at NGI, Science for Life Laboratory, Stockholm, Sweden, generating short (2x100 bp) paired reads. Raw sequences were base called using CASAVA RTA 1.18 (http://support.illumina.com/sequencing/sequencing_software/casava.htm) followed by removal of adapter sequences using Trimmomatic (Bolger, et al., 2014). Remaining read pairs were aligned to the Hg19 human reference genome using BWA (Li and Durbin, 2009) and discordant read mapping was called using TIDDIT (Eisfeldt, et al., 2017) as described in the previous section (*3.2.2.1 30X PCR-free paired-end sequencing*).

3.2.2.3 Linked-read sequencing

Linked-read sequencing was performed on two samples in **paper IV** and one sample in **paper V**, using 10X Genomics Chromium WGS protocol at NGI, Science for Life Laboratory, Stockholm, Sweden. Input DNA is kept as intact as possible to allow for linked reads spanning large genomic areas, and the long DNA fragments are referred to as molecules. Molecules are separated into oil droplets and a unique barcode is attached to the DNA in multiple in each droplet, followed by fragmentation of molecules into ~300 bp

fragments before sequencing on an Illumina sequencer. The linking of tagged reads is performed after sequencing, when each barcode is matched together with other reads containing the same barcode, hence belonging to the same original DNA molecule. The library was prepared using the 10X Chromium controller and sequencing was performed on an Illumina HiSeq Xten platform. Resulting data was analyzed using 10X Genomics-provided pipelines called Long Ranger and Supernova *de novo* assembler.

3.2.2.4 Validation of WGS results

All breakpoints where breakpoint junctions could be precisely mapped were confirmed with PCR and Sanger sequencing. Breakpoint PCR was performed using Phusion High-Fidelity DNA Polymerase (ThermoFisher Scientific) and sequenced using standard Sanger sequencing protocols. Sequences were aligned using BLAT (UCSC Genome Browser) (Kent, 2002) and visualized in CodonCode Aligner (CodonCode Corp., Dedham, MA, USA) (Figure 8). All cryptic CNVs in the breakpoints were confirmed using aCGH with the same protocols as described in section *3.2.1.2 Array comparative genomic hybridization (aCGH)*.

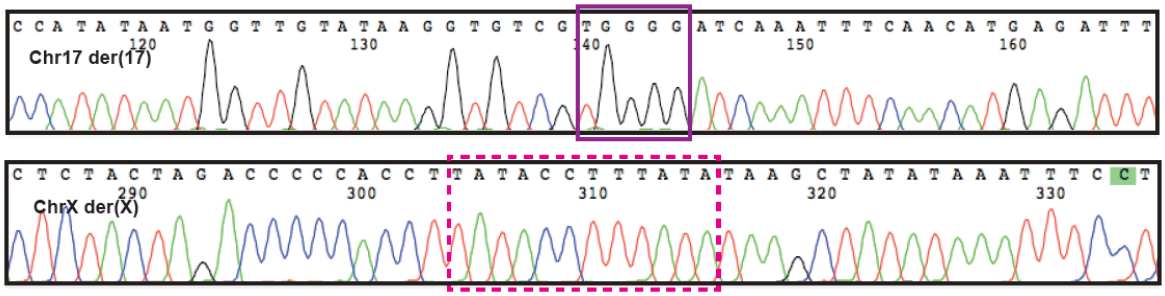

Figure 8. Sanger traces visualized in CodonCode Aligner software. *Above example is a balanced translocation between chromosomes 17 and X, with a 5 nt microhomology in the der(17) breakpoint junction (purple box) and a 12 nt insertion in the der(X) breakpoint junction (pink dotted line box). Picture is adapted from* ***Paper I****.*

3.2.3 Zebrafish studies

Functional studies in model organisms are crucial for correct interpretation of variants in novel disease genes or novel types of variants in known disease genes. Many genes are conserved across species, enabling *in vivo* studies of the pathophysiology of genetic aberrations in humans. Humans share over 98% of the genomic sequence with chimpanzee (*Pan troglodytes*) (Chimpanzee and Analysis, 2005), ~92% with mouse (*Mus musculus*) (Mouse Genome Sequencing, et al., 2002) and ~70% with zebrafish (*Danio rerio*) (Howe, et al., 2013). Zebrafish is a very popular model organism for early developmental studies, mainly because the fertilization and development occur *ex utero*. In addition, the fertilized eggs are transparent during the first week of development and therefore embryogenesis can easily be monitored. The majority of organs and tissues in humans are also present in zebrafish, for example vascular system, liver, brain and skeletal/cartilage tissues (Hammarsjo, et al., 2017; Hofmeister, et al., 2015; Hofmeister, et al., 2018; Laurell, et al., 2014; Song, et al., 2016; Wilkinson, et al., 2014). For quick screening of the impact of gene variants in the zebrafish model, gene expression can be easily modulated using knockdown or overexpression techniques. Injection of antisense oligonucleotides (morpholinos) results in gene knockdown, and injection of human mRNA (wild-type or with genetic variants found in patients) into the newly fertilized zebrafish eggs results in gene overexpression. The discovery of gene editing tools such as the CRISPR/Cas9 (and similar techniques developed in the previous years) and their application in model organisms has revolutionized the way researchers assess variants found in patients. This technique allows for stable mutagenesis of the study model carrying specific variants of interest, and mutations are passed on through the germline (Albadri, et al., 2017; Li, et al., 2016).

In **paper II**, we used a zebrafish model to assess the clinical impact of an intragenic *IFT81* duplication. The patient presented with less severe clinical symptoms than previously published cases with loss of IFT81 protein, and Western blot analysis suggested that a shorter isoform of the transcript was unaffected by the duplication. Hence, we wanted to test whether a truncated transcript of human *IFT81* could rescue the phenotype of an *ift81* knockdown. For this, we used morpholino knockdown to reduce endogenous *ift81* expression in combination with overexpression of human wildtype (wt) and truncated *IFT81* mRNA. Adult zebrafish were maintained on a 14 h day/10 h night cycle at Karolinska Institutet zebrafish core facility and embryos were produced by mass spawning. Injection of morpholinos and mRNA was performed at the 1-cell stage and embryos were maintained at 25°C until phenotype assessment at the 8-10 somite stage (~12-14 hours post fertilization).

4 RESULTS AND DISCUSSION

4.1 PAPER I

In **paper I**, we characterized 23 cytogenetically balanced translocations using low-coverage (3X) MP WGS and subsequent Sanger sequencing, hypothesizing that reciprocal translocations may be mediated by other mechanisms than non-homologous end-joining (NHEJ) or non-allelic homologous recombination (NAHR). In total, 46 breakpoint junctions from 22 carriers (one individual carried two separate reciprocal translocations) were characterized. Out of the total number of included cases, eight were reported clinically unaffected (36.4%) and 14 had an affected phenotype (63.6%).

Analysis of the breakpoint junctions revealed that protein-coding genes were disrupted in 48% of the breakpoints, and to the same extent in the clinically unaffected cohort as in the clinically affected cohort. However, in the clinically unaffected cohort the disrupted known disease genes were all recessive (*COG7, ALMS1, LARGE, OCA2*), and in the clinically affected cohort we identified three disrupted known dominant intellectual disability genes (*CTNND2, EXOC6B, GRIN2B*). Hence, three individuals received a molecular genetic diagnosis as a result of the study. It is also important to note that the four clinically unaffected carriers are heterozygous carriers of that recessive disease, in addition to being more prone to have fertility problems due to malsegregation of the rearranged chromosomes.

Five candidate genes for neurocognitive disabilities were identified (*SVOPL, SUSD1, TOX, NCALD, SLC4A10*), as well as two candidate genes for Tourette syndrome/tics (*LYPD6, GPC5*). In addition, a *de novo* 11.4 Mb deletion on chromosome 1p31, affecting 37 protein-coding genes and classified as pathogenic, was found in one individual.

Finally, detailed characterization of the 46 breakpoint junctions revealed typical mutational signatures of replication-based repair mechanisms such as FoSTeS/MMBIR in a substantial fraction of the translocations (17.4%, n=4). Such mutational signatures are typically microhomology and templated insertions, as well as rare SNVs in close proximity of the junction(s) (Figure 9). The findings indicated that FoSTeS/MMBIR is likely to be responsible for the formation of balanced reciprocal translocations to a larger extent than previously reported, possibly as high as 17% of interchromosomal translocation events.

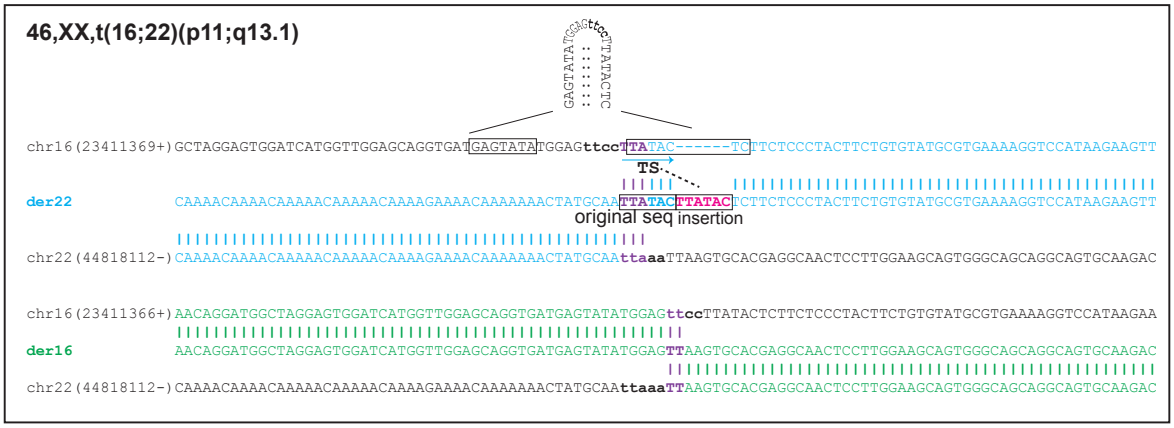

Figure 9. *Example of a reciprocal cytogenetically balanced translocation from* ***paper I*** *with mutational signatures of replication based repair mechanisms in the junctions of der22.*

None of the reported junctions were mediated by LCRs or other repeat elements. One translocation had both breakpoints within *Alu* elements but only a very short microhomology of 3 nt was present in both breakpoint junctions, indicating that no fusion *Alu* element was formed. However, a 1,579 bp deletion present 121 bp upstream of the breakpoint junction of one derivative chromosome was identified, generating a fusion *Alu* (*AluSx3-AluY*). These findings are consistent with a template-switching event, causing both the small deletion and the reciprocal translocation (Figure 10).

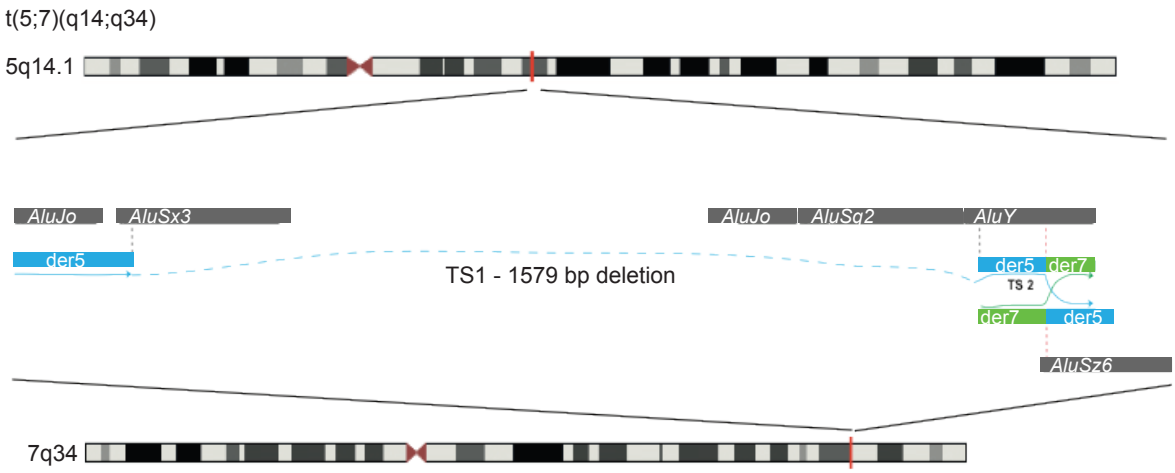

Figure 10. Schematic overview of 5q14.1 and 7q34 regions. *The derivative chromosome 5 (der5) is shown in blue and derivative chromosome 7 (der7) is shown in green. Alu elements are shown as dark grey boxes. The small deletion upstream of the breakpoint junction on der5 is shown as a dashed blue line. Both deletion breakpoints as well as the reciprocal translocation breakpoints are located within Alu elements. The figure is adapted and modified from* ***paper I****. TS; template-switching*

4.2 PAPER II

In **paper II**, we first performed targeted CNV screening in two individuals with distinct skeletal dysplasias (Patient 1: Jeune syndrome (MIM:208500), Patient 2: multiple epiphyseal dysplasia (MED) type 5 (MIM:607078)) and identified intragenic, exonic duplications in *IFT81* (Patient 1) and *MATN3* (Patient 2).

Both patients had previously been investigated with exome sequencing with a negative result. To assess whether small or large CNVs could explain the clinical phenotype, a customized aCGH was used targeting 1,989 genes previously implicated in ciliopathies, skeletal dysplasias, intellectual disability and congenital malformation syndromes. The custom aCGH analysis identified intragenic duplications in *IFT81* (12.4 kb) and *MATN3* (10.4 kb). The practical resolution of the regular clinical aCGH used at the Karolinska University Hospital is 25-50 kb, and both duplications would hence have been missed without the custom aCGH design. Both *IFT81* and *MATN3* had previously been implicated in skeletal dysplasias (*IFT81*; Jeune syndrome, *MATN3*; multiple epiphyseal dysplasia type V) concordant with the phenotypes of the included individuals.

Follow-up studies with short-read WGS and breakpoint PCR showed that both duplications were in tandem orientation and most likely disrupted the open reading frame (ORF) of the canonical transcript of both genes (Figure 11). In addition, we discovered that the duplication in *IFT81* was present in four copies (homozygous) and segregation analysis of the healthy, non-related parents revealed that both were heterozygous carriers of the rare duplication.

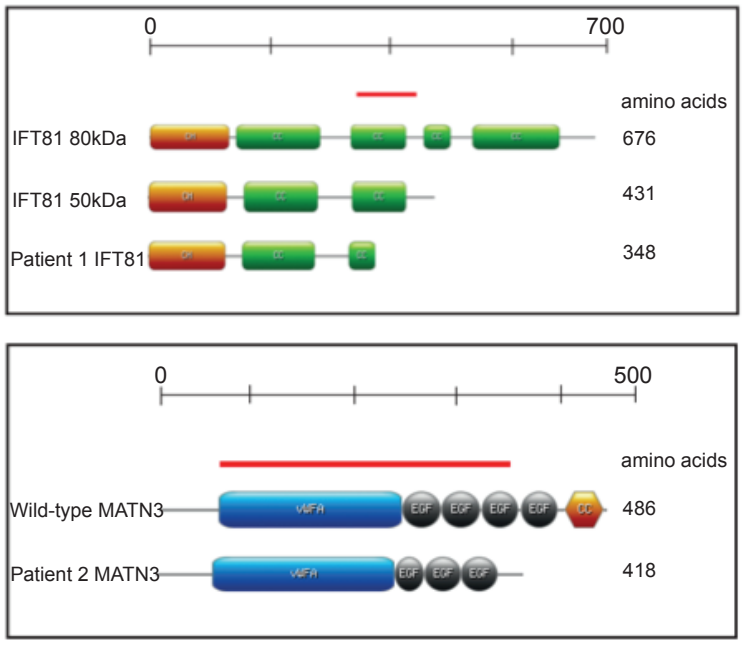

Figure 11. Intragenic duplications predicted to cause disruption of open reading frames and truncated proteins in both patients. *Both duplications were predicted to cause truncated proteins of both IFT81 and MATN3, possibly leading to nonsense-mediated decay. Picture adapted and modified from* ***paper II****.*

Hypothesizing that the parents of Patient 1, heterozygous for the same rare duplication in *IFT81*, shared a common ancestor we performed a loss of heterozygosity (LOH) analysis of the WGS data in the patient. The results showed that three small regions of the patient's genome showed LOH. One of the regions was a 4.8 Mb region on chromosome 12, which included the *IFT81* locus. The regions displaying LOH were however so small that the sample from Patient 1 clustered with the known non-related samples in a follow-up analysis (Figure 12).

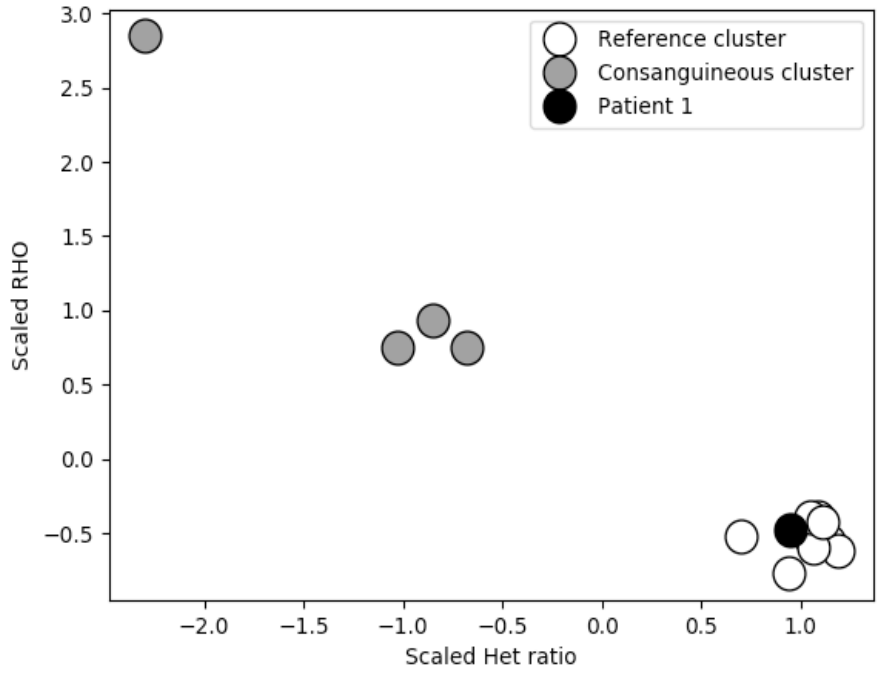

Figure 12. Analysis of loss of heterozygosity (LOH) in 14 samples. *WGS data from Patient 1 carrying the homozygous IFT81 duplication was analyzed for LOH and compared to four known consanguineous samples and nine known unrelated samples. It was found that Patient 1 (black) clustered with the unrelated samples, indicating that the common ancestor of the parents is several generations back. Figure adapted from* ***paper II****. RHO; regions of homozygosity*

Western blot analysis with protein extracted from fibroblasts from Patient 1 did not detect any wild-type full-length IFT81 protein (80 kDa), but only a shorter IFT81 isoform (50 kDa) that was also present in the healthy control fibroblasts. Complementary zebrafish studies suggested that even though the duplication results in loss of the full-length IFT81, remaining expression of a shorter IFT81 isoform is able to partially compensate the function of the full-length isoform, resulting in the milder form of Jeune syndrome seen in Patient 1.

Finally, split reads from the WGS data and breakpoint PCR showed that both duplications likely were *Alu-Alu* mediated with both breakpoints located within *Alu* elements and high sequence homology in the junctions.

4.3 PAPER III

In **paper III**, the proband was referred for genetic investigation due to developmental delay and mildly dysmorphic facial features and clinical aCGH identified a large duplication of approximately 1 Mb at chromosome 5p13. The duplication involved the *NIPBL* gene and duplications overlapping this specific locus cause the 5p13 duplication syndrome (MIM:613174).

During segregation analysis of the duplication it was found that the healthy mother carried two small deletions, flanking the region that was duplicated in the proband, which prompted for further analysis with WGS. It was found that in fact, the mother carried a total of five distinct deletions, sized between 20 kb and 100 kb, of which four were part of a chromosomal rearrangement with characteristics suggesting chromothripsis. In addition to the two small deletions flanking the duplication found in the proband a third deletion, sized only 20 kb and not detectable with regular aCGH (resolution ~50 kb), was present. Furthermore, the large duplication first identified in the proband was in fact two duplications of approximately 500 kb each. Finally, resolving the structure of the chromothriptic chromosome in the mother also revealed the origin of the disease-causing rearrangement in the proband; the derivative chromosome harboring the *NIPL* duplication had formed through unequal crossing-over during meiosis (Figure 13).

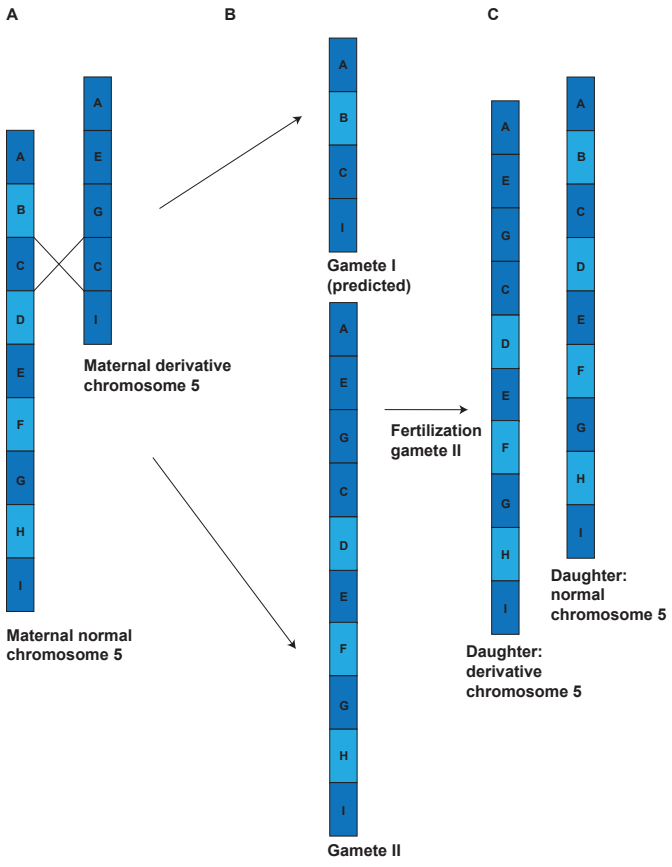

Figure 13. Unequal crossing-over during meiosis of a benign derivative chromosome 5 causes pathogenic duplications. *The benign derivative chromosome 5 in the mother, missing segments B, D, F and H* ***(A)*** *produced two unbalanced gametes* ***(B)****, where gamete II had rescued all deletions except segment B and gained extra copies of segments E and G. Fertilization of gamete II with the paternal normal chromosome 5 caused duplications of segments E and G* ***(C)****.*

Figure adapted and modified from ***paper III****.*

4.4 PAPER IV

In **paper IV** we aimed to characterize 16 unique (24 in total) cytogenetically detected chromosomal inversions in detail, in order to investigate the mechanism of formation and genotype-phenotype correlations in the clinically affected individuals. Out of the total of 16 unique inversions, we were able to characterize 11 (69%) at the nucleotide level. Among these 11 inversions, we found that two seemingly recurrent inversions actually were identical by descent with identical breakpoint junctions showing little to no microhomology. To further analyze those two inversions, we performed haplotype analysis on the WGS data, hypothesizing that the carriers would share haplotypes on the affected chromosomes (chromosome 10 and chromosome 12, respectively). The analysis showed that all individuals carrying the same inversion (inv(10)(p11.2q13) or inv(12)(p11.2q21)) shared both common and more rare haplotypes compared to 13 unrelated individuals of Swedish descent (p-values $2.8e^{-07}$ for inv(10) and $1.8e^{-08}$ for inv(12)).

Next, we found that two of the studied inversions were not copy number neutral; one, inv(X)(p22.31q28), had duplications flanking the inversion (DUP-INV-DUP) (Figure 14), and the other, inv(12)(q11.2q24.1) (Figure 15), involved both duplications and a deletion (DUP-INV-DEL-DUP) in addition to the inverted segment. Both of these complex inversions showed mutational signatures in the breakpoint junctions consistent with MMBIR, and the inv(X) was mediated by *Alu* elements.

Figure 14. A seemingly balanced inversion on chromosome X showed additional complexity in both breakpoints. *The inversion had been mediated by Alu elements, and the duplications had been formed concomitantly to the inversion. The distal breakpoints made the inversion prone to produce unbalanced gametes and segregation studies of the family identified two unbalanced recombinant chromosomes (rec(X)), with deletion of segment A (9.4 Mb) and duplications of segments D (58 kb) and E (1.8 Mb).*

Phasing the complex inv(X) rearrangement with WGS data from linked-read sequencing (10X Genomics Chromium) showed that both duplications originated from the same allele as the inversion and hence had been formed concomitantly with the inversion. The inversion and both duplications were mediated by *Alu* elements and fusion *Alus* were formed in both inversion breakpoint junctions. The breakpoints on Xq28 involved the Opsin gene/*TEX28* region that previously has been implicated in the formation of *MECP2* duplications and hence is known to be prone to rearrangements (Carvalho, et al., 2009). Finally, the family segregation studies revealed that two individuals in the family from two generations had unbalanced recombinant chromosomes resulting from inv(X): one adult female presenting with typical skeletal manifestations consistent with *SHOX* deletions and Leri-Weill dyschondrosteosis (MIM:127300), and one male with severe malformations who died *in utero*.

The inversion on chromosome 12 involved a small deletion (3.8 kb) and two small duplications (7.9 kb and 25 kb, originating from 12p and 12q, respectively). The final structure of the chromosome was predicted using the WGS data and confirmed using droplet digital PCR, showing that the B-E junction was indeed present twice. No genes were however affected by either of the CNVs or the inversion breakpoints.

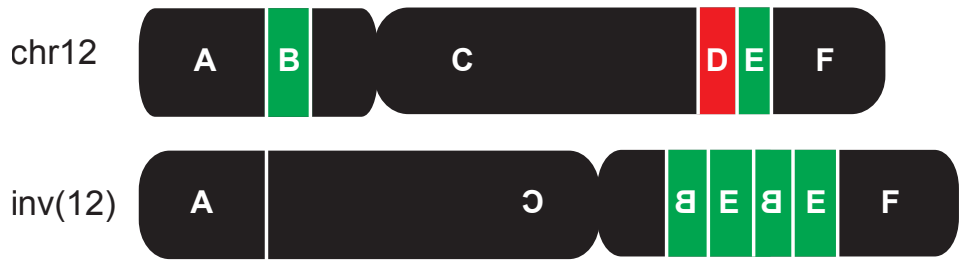

Figure 15. A seemingly balanced pericentric inversion on chromosome 12 had small imbalances in the breakpoints. *The predicted structure was determined by the WGS data and digital droplet PCR further confirmed that the B-E junction was indeed present twice. Figure is adapted and modified from* ***paper IV****.*

In conclusion, none of the inversions that were resolved at the nucleotide level (11/16, 69%) were mediated by ectopic recombination between inverted repeats, but instead most of the breakpoint junctions were simple and consistent with formation mechanisms commonly associated with reciprocal translocations, such as NHEJ. Two of the seemingly balanced inversions were not copy number neutral and showed mutational signatures consistent with replication errors and MMBIR. Finally, we showed that high-coverage short-read WGS detects a substantial fraction of chromosomal inversions with nucleotide resolution.

4.5 PAPER V

In paper V, we performed WGS on 21 individuals carrying multiple CNVs on the same chromosome arm aiming to characterize the breakpoint junctions (BPJs) in detail and delineate the structure of the rearranged chromosomes. A total of 83 BPJs were identified and all rearrangements were classified according to the patterns observed: deletions-only (n = 8), duplications-only (n = 7) and deletions-and-duplications (n = 6). In the deletions-only group, we observed additional structural rearrangements and BPJ characteristics typical to chromothripsis, while the duplications-only and deletions-and-duplications groups demonstrated mostly interspersed duplications and BPJs with microhomology. Two rearrangements were repetitive element-mediated (*Alu* and LINE, respectively) (Figure 17), and two rearrangements were found to be identical (2p25.3 clustered CNVs) at nucleotide level.

Detailed characterization of the rearranged chromosomes revealed that multiple cellular mechanisms are likely to be involved in the formation of clustered CNVs, such as breakage-fusion bridge cycles together with haltered formation of a ring chromosome, chromothripsis, chromoanagenesis, as well as at least two cell machineries operating simultaneously. The analysis also added further evidence for chromoanagenesis mechanisms underlying the formation of both "simple" and highly complex chromosomal rearrangements.

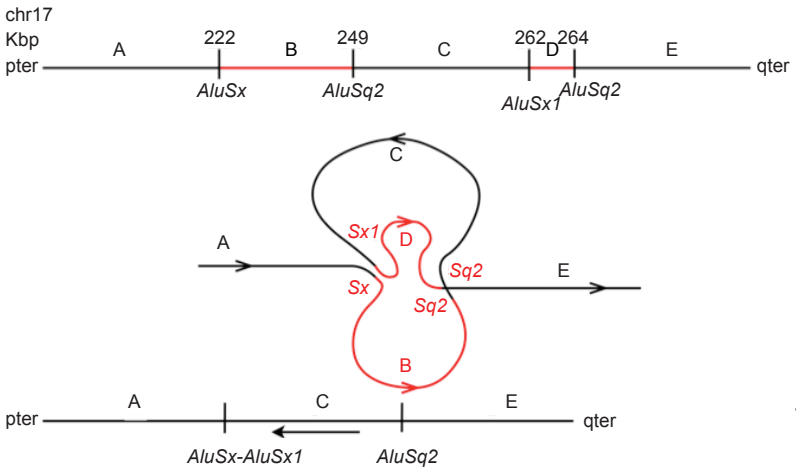

Figure 17. Schematic picture of Alu-Alu mediated rearrangement. *Case P2109_123 from* ***paper V*** *harbored a DEL-INV-DEL rearrangement (copy number state is indicated in black for normal copy number and red for copy number loss, inverted segments are indicated with an arrow). Analysis of the breakpoint junctions revealed an Alu-Alu fusion in junction A-C and all breakpoints were located within Alu elements (AluSx, AluSq2 and AluSx1). The figure is adapted and modified from* ***paper V****.*

5 CONCLUDING REMARKS

All studies included in this thesis have focused on detailed characterization of structural genomic variants to understand their role in human disease as well as underlying mechanisms of formation, in the hope to identify new or previously known disease genes that could explain the clinical symptoms and understand how, when, and why the rearrangements occurred. The following conclusions can be drawn from these studies:

- Whole genome sequencing of both seemingly balanced and complex SVs will add valuable information such as positional information, especially for duplications, and orientation of copy number neutral genomic segments. In a single experiment, it is possible to resolve the complete structure of many chromosomal rearrangements and identify cryptic aberrations not detected on microarray, such as inversions, translocations and small imbalances in the breakpoints. This added information is highly clinically relevant in many cases, and hence WGS is an important complementary method to aCGH, which is still the first screening method for SVs in clinical diagnostics.
- Reciprocal translocations are likely more commonly mediated by errors during DNA replication than what has previously been described and mechanisms involving template switiching might contribute to the formation of up to 17% of reciprocal translocations.
- Small intragenic duplications are rare, but important, human disease causing alleles. Targeted microarray analysis or whole-genome sequencing may be used to identify such small structural variants in a screening setting.
- Complex genomic rearrangements are commonly classified according to the mutational signatures in the breakpoint junctions, as well as whether the rearrangements involve deletions, duplications, or both. However, it is important to also perform detailed analysis of the parents that may carry a rearrangement that could have promoted the formation of the disease-causing rearrangement.
- Non-allelic homologous recombination (NAHR) is likely not the major mechanism underlying the formation of cytogenetically detected chromosomal inversions. Instead, our data suggests that most inversions show mutational signatures consistent with non-homologous end-joining, similar to what is seen in reciprocal translocations.

6 FUTURE PERSPECTIVES

Today, many of the monogenic disease-causing genes have been identified. Despite this fact, it is still difficult to prove that a particular gene is affected when no coding sequence variants are identified. *De novo* structural variants are quite common findings in developmental disorders, but assigning function to non-coding variation is and will continue to be a major challenge in human genetics. Additionally, to identify the genetic cause of phenotypes such as isolated intellectual disability and autism is still a demanding task. As mentioned previously, the numbers of mildly affected balanced chromosomal aberration carriers are unknown, and we most likely will continue to discover new candidate disease genes by looking thoroughly at apparently balanced chromosomal rearrangements, both with directly affected genes but also with position effects caused by the physical change of the 3D structure of the genome.

In addition, the continuous investigation of the underlying mechanisms of SV formation will help understanding the biological mechanisms underlying chromosomal rearrangements and using whole genome sequencing for exact pinpointing of breakpoint junctions provides the potential of discovering new biological mechanisms in genome stability and instability.

As WGS continues to be implemented in the clinical setting, we will continue to find new SVs that would not have been detected using cytogenetic tools. The clinical impact of these variants needs to be thoroughly studied, and for correct interpretation positional information, breakpoint junction architecture and resolution at nucleotide level is needed. In the clinical setting, it is of high importance to be able to filter out the normal variation of the genome from the rare and potentially pathogenic variants. To use WGS as a screening method for SVs, the databases of normal variation will have to be larger and include more populations, and bioinformatics tools to handle the massive amounts of data need to be optimized.

In conclusion, by understanding the mechanisms underlying structural variants formation we will also provide clues to how such events re-organize the human 3D genome and change expression of disease genes, with or without interfering with the coding sequence. In addition, WGS data will provide with clinically relevant information such as positional information, adding valuable information needed for clinical evaluation of the variants.

7 SAMMANFATTNING PÅ SVENSKA

En strukturell kromosomavvikelse innebär att den fysiska strukturen på en kromosom har förändrats på något sätt. Kopietalsavvikelser, så kallade *copy number variants* (CNVs) innebär att delar av kromosomen fattas (deletioner) eller har kopierats upp fler gånger än normalt (duplikationer/triplikationer). Eftersom dessa kromosomavvikelser innebär att genetiskt material har tillkommit eller fattas, kallas de obalanserade. Andra strukturella kromosomavvikelser kan innefatta fler än en kromosom, där delar av två eller fler kromosomer har gått av och bytt plats (translokationer), att en och samma kromosom gått av på två ställen och kromosommaterialet mellan brottspunkterna vridit sig 180° (inversion) eller att en bit av en kromosom brutit sig loss och satt sig på ett annat ställe (insertion). Dessa typer av kromosomavvikelser genererar vanligtvis ingen tillkomst eller förlust av genetiskt material och kallas därför balanserade. Riktigt komplexa kromosomavvikelser innehåller flertalet brottspunkter, ibland över hundra, och kan innefatta både obalanserade och balanserade strukturella varianter.

Bärare av balanserade strukturella kromosomavvikelser är ofta helt friska och rearrangemangen av kromosomerna upptäcks ofta först vid kromosomutredning på grund av infertilitet eller upprepade spontana aborter. Dessa beror ofta på att det balanserade kromosomrearrangemanget kan dela sig på flera olika sätt i könscellerna och på så sätt ibland orsaka obalanserade kromosomavvikelser hos fostret. I mycket ovanliga fall orsakar ändå en till synes balanserad kromosomavvikelse medicinska problem på grund av att viktiga arvsanlag (gener) ligger precis i brottspunkterna. Dessa har tidigare inte kunnat upptäckas på grund av för låg upplösning på metoderna som använts, men numera kan vi ofta hitta brottspunkterna på kromosomavvikelser ner på basparsnivå. Metoden som möjliggör detta kallas för helgenomsekvensering och innebär att vi genom ett enda experiment kan sekvensera hela arvsmassan (cirka 3 miljarder baspar innehållande cirka 20-25 000 gener) och pussla ihop allt genetiskt material för att förstå hur kromosomavvikelser sitter ihop och hur de kan ha uppstått.

Det finns ett flertal teorier kring uppkomstmekanismerna för kromosomavvikelser och de olika mekanismerna har distinkta ”signaturer” som kan avslöjas av sekvenserna kring brottspunkterna och där kromosombitarna limmats ihop. Ett flertal kända kromosomavvikelser uppstår gång på gång hos obesläktade individer, och det beror på så kallade repetitiva sekvenser. Mindre än 2% av allt DNA hos människan kodar för protein,

resten är så kallat ”ickekodande” genetiskt material som vi inte vet så mycket om ännu. Detta genetiska material innehåller stora delar sekvens som upprepas flera gånger och alltså finns på flera platser i genomet. När dessa sekvenser kommer för nära varandra, kan maskineriet som kopierar och reparerar DNA blanda ihop sekvenserna och kopiera/reparera DNA på ett felaktigt sätt. Detta kan då leda till att bitar av DNA förloras eller kopieras upp för många gånger, vilket kan orsaka sjukdomar ifall gener som är känsliga för kopietalsavvikelser är involverade.

Komplexa kromosomavvikelser med många brottspunkter kan uppstå genom flertalet olika mekanismer, där de två vanligaste innebär antingen en ”nybildning” av kromosomen där maskineriet för kopiering och reparation av DNA gör upprepade misstag (chromoanasynthesis), eller en ”kromosomkatastrof” där kromosomen av olika anledningar i det närmaste pulveriseras och därefter pusslas samman i en till synes helt slumpmässig ordning (chromothripsis).

Att förstå sambandet mellan DNA-struktur och hur kromosomavvikelser uppstår är helt avgörande för att kunna ge personer och familjer med kromosomavvikelser rätt vägledning och hjälp. Att få en diagnos har ofta stor inverkan på patientens och den närmaste familjens situation genom möjlighet till anlagstestning och fosterdiagnostik. Om vi förstår hur, var, när och varför kromosomavvikelser uppstår, kan vi även ge korrekt vägledning för exempelvis upprepningsrisk vid nya graviditeter. I vissa fall kan även rätt diagnos ge möjlighet till medicinsk behandling för att lindra eller bota symptom.

Syftet med denna avhandling har varit att försöka förstå de bakomliggande uppkomstmekanismer som orsakar kromosomavvikelser samt identifiera både nya och tidigare kända sjukdomsorsakande gener. Sammanfattningsvis visar resultaten att misstag under DNA-replikation troligtvis är en vanligare underliggande uppkomstmekanism för balanserade kromosomavvikelser än vad som tidigare rapporterats. Avhandlingen visar även tydligt att helgenomsekvensering är en mycket viktig metod för att kunna definiera och analysera kromosomavvikelser på djupet. På lång sikt kan fynden förhoppningsvis vara till hjälp för att ge fler patienter rätt diagnos och fler familjer hjälp med anlagstest, fosterdiagnostik och vägledning.

8 ACKNOWLEDGEMENTS

I am deeply grateful to all patients, their families and their treating physicians who made this research possible.

My main supervisor, associate professor **Anna Lindstrand**, for all your support, guidance, motivation, positivity and for truly believing in me from day one. I am forever grateful for your trust in me when first taking me in as a bachelor student and for always being there for me whenever I needed you. I feel confident in what I am capable of and I owe you for that. I could not have wished for a better supervisor!

Daniel Nilsson, co-supervisor, for your support and guidance through the bioinformatics and your positivity.

Magnus Nordenskjöld, co-supervisor, for first introducing me to the world of genetics, welcoming me into the research group back in 2013 and supporting me when I wanted to start the doctoral studies.

Elisabeth Syk Lundberg, co-supervisor, for your enthusiastic and excellent guidance in the exciting field of chromosomes.

Johanna Lundin, co-supervisor, for sharing your cytogenetics expertise and for the enthusiasm you always show when I ask for your help.

Former and present members of the Rare Diseases research group, Clinical Genetics research group, collaborators, colleagues and fellow researchers: **Agne Liedén, Agneta Nordenskjöld, Alexandra Löfstedt, Alice Costantini, Alisa Förster, Anders Kämpe, Ann Nordgren, Anna Flögel, Anna-Lena Kastman, Bianca Tesi, Britt-Marie Anderlid, Cecilia Arthur, Christa Costa, Christina Nyström, Christopher Grochowski, Claudia Carvalho Fonseca, Diego Cortese, Emeli Pontén, Erik Iwarsson, Giedre Grigelioniene, Helena Malmgren, Hero Nikdin, Håkan Thonberg, Ingegerd Ivanov Öfverholm, Isabel Tapia Paez, Jessica Alm, Johanna Winberg, Josephine Wincent, Karin Salehi Karlslätt, Karin Wallander, Kicki Lagerstedt, Malin Kvarnung, Mastoureh Shahsavani, Miriam Armenio, Måns Magnusson, Nadja Pekkola Pacheco, Nina Jäntti, Outi Mäkitie, Peter Gustavsson, Raquel Vaz, Samina Asad, Sara Dahl, Sofia Frisk, Tobias Laurell, Vasilios Zachariadis** and **Wolfgang Hofmeister**.

Special thanks to **Ellika Sahlin**, for being a really great friend and the funniest person I know! These years would not have been the same without you!

Thank you **Fulya Taylan** and **Benedicte Bang** for being great friends and gym-partners, **Jesper Eisfeldt** for your infinite patience and for always looking at things from the positive side and **Anna Hammarsjö**, for fun talks and great collaborations. →

Researchers, clinicians and co-authors that have contributed to the projects in this thesis, without you I would not be here and for all your hard work, I am forever grateful.

The administrative staff at MMK, with special thanks to **Ann-Britt Wikström**.

Thank you to **all of my friends** who have supported me throughout the years!

Special thanks to **Hannah Schwartz** for painting the amazing art that is on the cover of this book, **Ulrika Lidbjörk** for being the best friend one could ever ask for even when you are on the other side of the planet, and **Katarina Sjöö** for always having my back and supporting me no matter what.

My amazing family: My parents **Barbro** and **Thomas**, you have always supported me in everything and I am forever grateful for that. My sister **Anna** who is the best sister one could ever ask for, I am so proud of you! Thank you all for always believing in me and never ever leaving my side.

My extra-family: **Irene**, **Claes-Göran**, **Malin**, **Jukka** and my little love **Alba**.

Marcus, the love of my life. You are everything to me!

9 REFERENCES

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res 21(6):974-84.

Albadri S, Del Bene F, Revenu C. 2017. Genome editing using CRISPR/Cas9-based knock-in approaches in zebrafish. Methods 121-122:77-85.

Aristidou C, Theodosiou A, Bak M, Mehrjouy MM, Constantinou E, Alexandrou A, Papaevripidou I, Christophidou-Anastasiadou V, Skordis N, Kitsiou-Tzeli S, et al. 2018. Position effect, cryptic complexity, and direct gene disruption as disease mechanisms in de novo apparently balanced translocation cases. PLoS One 13(10):e0205298.

Barbouti A, Stankiewicz P, Nusbaum C, Cuomo C, Cook A, Hoglund M, Johansson B, Hagemeijer A, Park SS, Mitelman F, et al. 2004. The breakpoint region of the most common isochromosome, i(17q), in human neoplasia is characterized by a complex genomic architecture with large, palindromic, low-copy repeats. Am J Hum Genet 74(1):1-10.

Blackburn EH. 1991. Structure and function of telomeres. Nature 350(6319):569-73.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114-20.

Boone PM, Liu P, Zhang F, Carvalho CM, Towne CF, Batish SD, Lupski JR. 2011. Alu-specific microhomology-mediated deletion of the final exon of *SPAST* in three unrelated subjects with hereditary spastic paraplegia. Genet Med 13(6):582-92.

Boone PM, Yuan B, Campbell IM, Scull JC, Withers MA, Baggett BC, Beck CR, Shaw CJ, Stankiewicz P, Moretti P, et al. 2014. The Alu-rich genomic architecture of *SPAST* predisposes to diverse and functionally distinct disease-associated CNV alleles. Am J Hum Genet 95(2):143-61.

Bramswig NC, Ludecke HJ, Pettersson M, Albrecht B, Bernier RA, Cremer K, Eichler EE, Falkenstein D, Gerdts J, Jansen S, et al. 2017. Identification of new *TRIP12* variants and detailed clinical evaluation of individuals with non-syndromic intellectual disability with or without autism. Hum Genet 136(2):179-192.

Carvalho CM, Lupski JR. 2016. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet 17(4):224-38.

Carvalho CM, Ramocki MB, Pehlivan D, Franco LM, Gonzaga-Jauregui C, Fang P, McCall A, Pivnick EK, Hines-Dowell S, Seaver LH, et al. 2011. Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. Nat Genet 43(11):1074-81.

Carvalho CM, Zhang F, Liu P, Patel A, Sahoo T, Bacino CA, Shaw C, Peacock S, Pursley A, Tavyev YJ, et al. 2009. Complex rearrangements in patients with duplications of MECP2 can occur by fork stalling and template switching. Hum Mol Genet 18(12):2188-203.

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Kallberg M, Cox AJ, Kruglyak S, Saunders CT. 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32(8):1220-2.

Chiang C, Jacobsen JC, Ernst C, Hanscom C, Heilbut A, Blumenthal I, Mills RE, Kirby A, Lindgren AM, Rudiger SR, et al. 2012. Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. Nat Genet 44(4):390-7, S1.

Chimpanzee S, Analysis C. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437(7055):69-87.
Currall BB, Chiang C, Talkowski ME, Morton CC. 2013. Mechanisms for Structural Variation in the Human Genome. Curr Genet Med Rep 1(2):81-90.
de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet 7(12):e1002384.
Eisfeldt J, Pettersson M, Vezzi F, Wincent J, Kaller M, Gruselius J, Nilsson D, Syk Lundberg E, Carvalho CMB, Lindstrand A. 2019. Comprehensive structural variation genome map of individuals carrying complex chromosomal rearrangements. PLoS Genet 15(2):e1007858.
Eisfeldt J, Vezzi F, Olason P, Nilsson D, Lindstrand A. 2017. TIDDIT, an efficient and comprehensive structural variant caller for massive parallel sequencing data. F1000Res 6:664.
English AC, Salerno WJ, Hampton OA, Gonzaga-Jauregui C, Ambreth S, Ritter DI, Beck CR, Davis CF, Dahdouli M, Ma S, et al. 2015. Assessing structural variation in a personal genome-towards a human reference diploid genome. BMC Genomics 16:286.
Finelli P, Pincelli AI, Russo S, Bonati MT, Recalcati MP, Masciadri M, Giardino D, Cavagnini F, Larizza L. 2007. Disruption of friend of *GATA 2* gene (FOG-2) by a de novo t(8;10) chromosomal translocation is associated with heart defects and gonadal dysgenesis. Clin Genet 71(3):195-204.
Fountain JW, Wallace MR, Bruce MA, Seizinger BR, Menon AG, Gusella JF, Michels VV, Schmidt MA, Dewald GW, Collins FS. 1989. Physical mapping of a translocation breakpoint in neurofibromatosis. Science 244(4908):1085-7.
Fruhmesser A, Blake J, Haberlandt E, Baying B, Raeder B, Runz H, Spreiz A, Fauth C, Benes V, Utermann G, et al. 2013. Disruption of *EXOC6B* in a patient with developmental delay, epilepsy, and a *de novo* balanced t(2;8) translocation. Eur J Hum Genet 21(10):1177-80.
Fukushi D, Yamada K, Suzuki K, Inaba M, Nomura N, Suzuki Y, Katoh K, Mizuno S, Wakamatsu N. 2018. Clinical and genetic characterization of a patient with *SOX5* haploinsufficiency caused by a *de novo* balanced reciprocal translocation. Gene 655:65-70.
Fullwood MJ, Wei CL, Liu ET, Ruan Y. 2009. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Res 19(4):521-32.
Funderburk SJ, Spence MA, Sparkes RS. 1977. Mental retardation associated with "balanced" chromosome rearrangements. Am J Hum Genet 29(2):136-41.
Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global reference for human genetic variation. Nature 526(7571):68-74.
Geoffroy V, Stoetzel C, Scheidecker S, Schaefer E, Perrault I, Bar S, Kroll A, Delbarre M, Antin M, Leuvrey AS, et al. 2018. Whole-genome sequencing in patients with ciliopathies uncovers a novel recurrent tandem duplication in *IFT140*. Hum Mutat 39(7):983-992.
Gimelli G, Pujana MA, Patricelli MG, Russo S, Giardino D, Larizza L, Cheung J, Armengol L, Schinzel A, Estivill X, et al. 2003. Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. Hum Mol Genet 12(8):849-58.
Gu S, Yuan B, Campbell IM, Beck CR, Carvalho CM, Nagamani SC, Erez A, Patel A, Bacino CA, Shaw CA, et al. 2015. *Alu*-mediated diverse and complex pathogenic copy-number variants within human chromosome 17 at p13.3. Hum Mol Genet 24(14):4061-77.

Halgren C, Nielsen NM, Nazaryan-Petersen L, Silahtaroglu A, Collins RL, Lowther C, Kjaergaard S, Frisch M, Kirchhoff M, Brondum-Nielsen K, et al. 2018. Risks and Recommendations in Prenatally Detected *De Novo* Balanced Chromosomal Rearrangements from Assessment of Long-Term Outcomes. Am J Hum Genet 102(6):1090-1103.

Hammarsjo A, Wang Z, Vaz R, Taylan F, Sedghi M, Girisha KM, Chitayat D, Neethukrishna K, Shannon P, Godoy R, et al. 2017. Novel *KIAA0753* mutations extend the phenotype of skeletal ciliopathies. Sci Rep 7(1):15585.

Hodge JC, Mitchell E, Pillalamarri V, Toler TL, Bartel F, Kearney HM, Zou YS, Tan WH, Hanscom C, Kirmani S, et al. 2014. Disruption of *MBD5* contributes to a spectrum of psychopathology and neurodevelopmental abnormalities. Mol Psychiatry 19(3):368-79.

Hofmeister W, Nilsson D, Topa A, Anderlid BM, Darki F, Matsson H, Tapia Paez I, Klingberg T, Samuelsson L, Wirta V, et al. 2015. *CTNND2*-a candidate gene for reading problems and mild intellectual disability. J Med Genet 52(2):111-22.

Hofmeister W, Pettersson M, Kurtoglu D, Armenio M, Eisfeldt J, Papadogiannakis N, Gustavsson P, Lindstrand A. 2018. Targeted copy number screening highlights an intragenic deletion of *WDR63* as the likely cause of human occipital encephalocele and abnormal CNS development in zebrafish. Hum Mutat 39(4):495-505.

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature 496(7446):498-503.

International Human Genome Sequencing C. 2004. Finishing the euchromatic sequence of the human genome. Nature 431(7011):931-45.

Jacobs PA, Browne C, Gregson N, Joyce C, White H. 1992. Estimates of the frequency of chromosome abnormalities detectable in unselected newborns using moderate levels of banding. J Med Genet 29(2):103-8.

Jacobs PS, Hassold, T. 1987. Chromosomal abnormalities: origin and etiology in abortions and live births. In: Vogel F, Sperling K, , eds. Human Genetics. Berlin: Springer-Verlag:233-244.

Kazazian HH, Jr. 2004. Mobile elements: drivers of genome evolution. Science 303(5664):1626-32.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. Genome Res 12(4):656-64.

Kloosterman WP, Guryev V, van Roosmalen M, Duran KJ, de Bruijn E, Bakker SC, Letteboer T, van Nesselrooij B, Hochstenbach R, Poot M, et al. 2011. Chromothripsis as a mechanism driving complex *de novo* structural rearrangements in the germline. Hum Mol Genet 20(10):1916-24.

Kloosterman WP, Tavakoli-Yaraki M, van Roosmalen MJ, van Binsbergen E, Renkens I, Duran K, Ballarati L, Vergult S, Giardino D, Hansson K, et al. 2012. Constitutional chromothripsis rearrangements involve clustered double-stranded DNA breaks and nonhomologous repair mechanisms. Cell Rep 1(6):648-55.

Krijger PH, de Laat W. 2016. Regulation of disease-associated gene expression in the 3D genome. Nat Rev Mol Cell Biol 17(12):771-782.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409(6822):860-921.

Laurell T, Nilsson D, Hofmeister W, Lindstrand A, Ahituv N, Vandermeer J, Amilon A, Anneren G, Arner M, Pettersson M, et al. 2014. Identification of three novel *FGF16* mutations in X-linked recessive fusion of the fourth and fifth metacarpals and possible correlation with heart disease. Mol Genet Genomic Med 2(5):402-11.

Lee JA, Carvalho CM, Lupski JR. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 131(7):1235-47.
Lettice LA, Daniels S, Sweeney E, Venkataraman S, Devenney PS, Gautier P, Morrison H, Fantes J, Hill RE, FitzPatrick DR. 2011. Enhancer-adoption as a mechanism of human developmental disease. Hum Mutat 32(12):1492-9.
Lettice LA, Horikoshi T, Heaney SJ, van Baren MJ, van der Linde HC, Breedveld GJ, Joosse M, Akarsu N, Oostra BA, Endo N, et al. 2002. Disruption of a long-range cis-acting regulator for *Shh* causes preaxial polydactyly. Proc Natl Acad Sci U S A 99(11):7548-53.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754-60.
Li M, Zhao L, Page-McCaw PS, Chen W. 2016. Zebrafish Genome Engineering Using the CRISPR-Cas9 System. Trends Genet 32(12):815-827.
Lieden A, Kvarnung M, Nilssson D, Sahlin E, Lundberg ES. 2014. Intragenic duplication--a novel causative mechanism for *SATB2*-associated syndrome. Am J Med Genet A 164A(12):3083-7.
Lindstrand A, Frangakis S, Carvalho CM, Richardson EB, McFadden KA, Willer JR, Pehlivan D, Liu P, Pediaditakis IL, Sabo A, et al. 2016. Copy-Number Variation Contributes to the Mutational Load of Bardet-Biedl Syndrome. Am J Hum Genet 99(2):318-36.
Lindstrand A, Malmgren H, Verri A, Benetti E, Eriksson M, Nordgren A, Anderlid BM, Golovleva I, Schoumans J, Blennow E. 2010. Molecular and clinical characterization of patients with overlapping 10p deletions. Am J Med Genet A 152A(5):1233-43.
Liu P, Carvalho CM, Hastings PJ, Lupski JR. 2012. Mechanisms for recurrent and complex human genomic rearrangements. Curr Opin Genet Dev 22(3):211-20.
Liu P, Erez A, Nagamani SC, Dhar SU, Kolodziejska KE, Dharmadhikari AV, Cooper ML, Wiszniewska J, Zhang F, Withers MA, et al. 2011. Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. Cell 146(6):889-903.
Lu H, Giordano F, Ning Z. 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics Proteomics Bioinformatics 14(5):265-279.
Lupianez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R, et al. 2015. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell 161(5):1012-1025.
Lupianez DG, Spielmann M, Mundlos S. 2016. Breaking TADs: How Alterations of Chromatin Domains Result in Disease. Trends Genet 32(4):225-237.
Lupski JR. 1998. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet 14(10):417-22.
Madan K, Nieuwint AW, van Bever Y. 1997. Recombination in a balanced complex translocation of a mother leading to a balanced reciprocal translocation in the child. Review of 60 cases of balanced complex translocations. Hum Genet 99(6):806-15.
Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J, Shago M, Moessner R, Pinto D, Ren Y, et al. 2008. Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet 82(2):477-88.
Masset H, Hestand MS, Van Esch H, Kleinfinger P, Plaisancie J, Afenjar A, Molignier R, Schluth-Bolard C, Sanlaville D, Vermeesch JR. 2016. A Distinct Class of Chromoanagenesis Events Characterized by Focal Copy Number Gains. Hum Mutat 37(7):661-8.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. Science 337(6099):1190-5.
McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. Genome Biol 17(1):122.
Metzker ML. 2010. Sequencing technologies - the next generation. Nat Rev Genet 11(1):31-46.
Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, et al. 2011. Mapping copy number variation by population-scale genome sequencing. Nature 470(7332):59-65.
Mitchell LA, Wang A, Stracquadanio G, Kuang Z, Wang X, Yang K, Richardson S, Martin JA, Zhao Y, Walker R, et al. 2017. Synthesis, debugging, and effects of synthetic chromosome consolidation: synVI and beyond. Science 355(6329).
Mouse Genome Sequencing C, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420(6915):520-62.
Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. 2003. Scanning human gene deserts for long-range enhancers. Science 302(5644):413.
Ordulu Z, Kammin T, Brand H, Pillalamarri V, Redin CE, Collins RL, Blumenthal I, Hanscom C, Pereira S, Bradley I, et al. 2016. Structural Chromosomal Rearrangements Require Nucleotide-Level Resolution: Lessons from Next-Generation Sequencing in Prenatal Diagnosis. Am J Hum Genet 99(5):1015-1033.
Osborne LR, Li M, Pober B, Chitayat D, Bodurtha J, Mandel A, Costa T, Grebe T, Cox S, Tsui LC, et al. 2001. A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. Nat Genet 29(3):321-5.
Ou Z, Stankiewicz P, Xia Z, Breman AM, Dawson B, Wiszniewska J, Szafranski P, Cooper ML, Rao M, Shao L, et al. 2011. Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes. Genome Res 21(1):33-46.
Pennisi E. 2010. Genomics. 1000 Genomes Project gives new map of genetic diversity. Science 330(6004):574-5.
Pentao L, Wise CA, Chinault AC, Patel PI, Lupski JR. 1992. Charcot-Marie-Tooth type 1A duplication appears to arise from recombination at repeat sequences flanking the 1.5 Mb monomer unit. Nat Genet 2(4):292-300.
Pinkel D, Straume T, Gray JW. 1986. Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. Proc Natl Acad Sci U S A 83(9):2934-8.
Redin C, Brand H, Collins RL, Kammin T, Mitchell E, Hodge JC, Hanscom C, Pillalamarri V, Seabra CM, Abbott MA, et al. 2017. The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. Nat Genet 49(1):36-45.
Reiter LT, Hastings PJ, Nelis E, De Jonghe P, Van Broeckhoven C, Lupski JR. 1998. Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple *HNPP* deletion patients. Am J Hum Genet 62(5):1023-33.
Rhoads A, Au KF. 2015. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics 13(5):278-89.
Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat Biotechnol 29(1):24-6.
Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. 2012. Linking disease associations with regulatory information in the human genome. Genome Res 22(9):1748-59.

Schluth-Bolard C, Labalme A, Cordier MP, Till M, Nadeau G, Tevissen H, Lesca G, Boutry-Kryza N, Rossignol S, Rocas D, et al. 2013. Breakpoint mapping by next generation sequencing reveals causative gene disruption in patients carrying apparently balanced chromosome rearrangements with intellectual deficiency and/or congenital malformations. J Med Genet 50(3):144-50.
Song X, Beck CR, Du R, Campbell IM, Coban-Akdemir Z, Gu S, Breman AM, Stankiewicz P, Ira G, Shaw CA, et al. 2018. Predicting human genes susceptible to genomic instability associated with *Alu*/*Alu*-mediated rearrangements. Genome Res.
Song Z, Zhang X, Jia S, Yelick PC, Zhao C. 2016. Zebrafish as a Model for Human Ciliopathies. J Genet Genomics 43(3):107-20.
Stankiewicz P, Lupski JR. 2002. Genome architecture, rearrangements and genomic disorders. Trends Genet 18(2):74-82.
Stankiewicz P, Shaw CJ, Dapper JD, Wakui K, Shaffer LG, Withers M, Elizondo L, Park SS, Lupski JR. 2003. Genome architecture catalyzes nonrecurrent chromosomal rearrangements. Am J Hum Genet 72(5):1101-16.
Stankiewicz P, Shaw CJ, Withers M, Inoue K, Lupski JR. 2004. Serial segmental duplications during primate evolution result in complex human genome architecture. Genome Res 14(11):2209-20.
Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, et al. 2015. An integrated map of structural variation in 2,504 human genomes. Nature 526(7571):75-81.
Talkowski ME, Ordulu Z, Pillalamarri V, Benson CB, Blumenthal I, Connolly S, Hanscom C, Hussain N, Pereira S, Picker J, et al. 2012. Clinical diagnosis by whole-genome sequencing of a prenatal sample. N Engl J Med 367(23):2226-32.
Tchinda J, Lee C. 2006. Detecting copy number variation in the human genome using comparative genomic hybridization. Biotechniques 41(4):385, 387, 389 passim.
Waldman AS, Liskay RM. 1988. Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology. Mol Cell Biol 8(12):5350-7.
Warburton D. 1980. Current techniques in chromosome analysis. Pediatr Clin North Am 27(4):753-69.
Warburton D. 1991. *De novo* balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: clinical significance and distribution of breakpoints. Am J Hum Genet 49(5):995-1013.
Wilkinson RN, Jopling C, van Eeden FJ. 2014. Zebrafish as a model of cardiac disease. Prog Mol Biol Transl Sci 124:65-91.
Yunis JJ, Sanchez O. 1973. G-banding and chromosome structure. Chromosoma 44(1):15-23.
Zarrei M, MacDonald JR, Merico D, Scherer SW. 2015. A copy number variation map of the human genome. Nat Rev Genet 16(3):172-83.
Zhang CZ, Spektor A, Cornils H, Francis JM, Jackson EK, Liu S, Meyerson M, Pellman D. 2015. Chromothripsis from DNA damage in micronuclei. Nature 522(7555):179-84.
Zhang F, Gu W, Hurles ME, Lupski JR. 2009a. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet 10:451-81.
Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, Lupski JR. 2009b. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nat Genet 41(7):849-53.